New Methods of Density Functional Optmization and *ab intio* Molecular Dynamics Convergence Testing Applied to Water

FACULTAD DE CIENCIAS
UNIVERSIDAD AUTÓNOMA DE MADRID

Michelle Fritz

Departamento de Física de la Materia Condensada y Nanotecnología
Universidad Autónoma de Madrid

This dissertation is submitted for the degree of
Doctorado en Física de la Materia Condensada y Nanotecnología

February 2016

Table of contents

List of figures **v**

List of tables **xiii**

1 Introduction **1**

1 Introducción **5**

2 Theory **8**

2.1 The Hartree-Fock approximation 8
2.2 Density functionals 11
2.3 Exchange and Correlation Functionals 14
2.3.1 The Exchange and Correlation hole 15
2.3.2 The Adiabatic Approximation 15
2.3.3 First rung: The Local Density Approximation 17
2.3.4 Second rung: Generalized-Gradient Approximation 17
2.3.5 Beyond the second rung 18
2.4 van der Waals interactions 19
2.5 Self-interaction errors in DFT 20
2.6 The SIESTA method 21
2.6.1 Pseudopotentials 22
2.6.2 Basis sets 23

3 Simulating Water with *Ab intio* Methods **25**

3.1 Liquid Water Interactions and Structure 25
3.2 The Importance of vdW Interactions, Self-interaction Correction, and Nuclear Quantum Effects in Water 28
3.2.1 vdW Interactions 28

3.2.2 Self-interaction Errors . . . 31
3.2.3 Nuclear Quantum Effects . . . 32
3.2.4 Conclusions . . . 33

4 The DPPS method and optimization of GGA exchange functionals for water 34
4.1 Methodology . . . 35
4.1.1 Data Projection onto Parameter Space . . . 35
4.1.2 Functional optimization . . . 36
4.1.3 Bayesian constraints . . . 39
4.2 Optimization of a GGA exchange functional for water . . . 40
4.2.1 Computational details . . . 41
4.2.2 Test: Does DPPS find the optimal GGA exchange functional? 47
4.2.3 Dataset details . . . 48
4.2.4 Optimized functionals . . . 52
4.2.5 1-body error corrections . . . 73
4.3 Conclusions . . . 77

5 The Thermodynamic Integration (TI) method for AIMD convergence testing 80
5.1 Convergence testing with AIMD simulations . . . 81
5.1.1 Property dependence on average simulation temperature . . 82
5.1.2 AIMD results . . . 84
5.2 The Thermodynamic Integration method . . . 86
5.2.1 Methodology . . . 86
5.2.2 Testing of the TI method . . . 88
5.2.3 Application of the TI method to convergence tests . . . 93
5.3 Conclusions . . . 94

6 Application of optimized functionals to gas and condensed phases of water 100
6.1 Total energy analysis of clusters . . . 101
6.1.1 Relative binding energies of hexamers . . . 101
6.1.2 Total energy analysis of water clusters from liquid snapshots 104
6.2 Ab initio Molecular Dynamics results . . . 107
6.2.1 Conclusions . . . 120

7 Conclusions **123**

7 Conclusiones **128**

Appendix A Constraints for the exchange-correlation energy **132**

References **134**

List of figures

3.1 Each water molecule, composed of one oxygen atom (red) and two hydrogen atoms (white) is part of the formation of a local tetrahedron in a H-bond network. The grey lines represent H-bond attractions between water molecules. 27

3.2 29

3.2 Taken from Ref. [46]. This figure shows the combined (a) and individual (b) effects of including vdW interactions, self interaction corrections, and nuclear quantum effects in liquid water simulations. In (a), the oxygen-oxygen RDF $g_{OO}(r)$ of liquid water obtained from an AIMD simulation at 300 K with PBE0+TS-vdW(SC), which takes into account all three effects, is shown to be in agreement with the experimental $g_{OO}(r)$ obtained from x-ray and scattering experiments [116, 18]. Here Q_{max} is the maximum value of Q, the change in the wavevector between the incident and scattered radiation, in the given scattering experiment. In (b), the $g_{OO}(r)$ of liquid water obtained from different DFT-based AIMD simulations are plotted. The arrows indicate the direction of the systematic shifts in the main RDF peak positions and intensities as the functionals are improved. . . . 30

4.1 The GGA exchange enhancement factors $F_x(s)$ of the LV (blue), PW86R (red), and LV-PW86R (green) exchange functionals. 42

4.2 Convergence test of 1-,2-, and 3-body energies with the $^{(p)}$dζ+p, dζ+p, tζ+dp, and $^{(s)}$qζ+dp basis sets. Plotted here are (a) the difference of the total energies (ΔE) with respect to qζ+dp and (b) the wall clock times of the monomer energy calculations for one SCF iteration with different bases. A 20 Å lattice constant and a 300 Ry mesh cutoff were used. The inset shows the values for $^{(s)}$qζ+dp. 43

4.3 Plotted here are (a) the errors of 1-,2-, and 3-body energies with respect to those calculated with a 50 Å lattice constant and (b) the calculation times for one SCF iteration of a monomer energy calculation with different lattice constants. A 300 Ry mesh cutoff and the $^{(s)}$qζ+dp basis set were used. 44

4.4 Plotted here are (a) the errors of 1-,2-, and 3-body energies with respect to a those calculated with a 500 Ry mesh cutoff and (b) times for one SCF iteration with different mesh cutoffs in the calculation of the energy of a water monomer. A 20 Å lattice constant and $^{(s)}$qζ+dp basis set were used. 45

4.5 PBE and PW86R exchange enhancement factors (blue and green lines, respectively) along with the GGA exchange enhancement factor values $F_{x\alpha}$ at interpolation points s_α, optimized without constraints to reproduce PW86R exchange energies of 198 water dimers (red asterisks). 49

4.6 Grid placement for symmetric water monomer geometries and MB-pol energy contours. The y-axis represents the H-O-H angle θ, and the x-axis represents the distance r between O and H. The contour values are at increments of 0.2 eV. 50

4.7 Grid placement for 330 water dimers. The y-axis represents the O-O distance r_{OO}, and the x-axis represents the dimer configuration index. The configurations represented by the indices are shown in the key on the bottom. An "H" represents a hydrogen atom and *L* represents a lone pair. The letters on each side of the dash represent the portion of the water monomer that is most directly facing the other monomer. For example, in the "H-HLL" configuration, the hydrogen atom (H) of the first monomer is facing the second molecule, while the second monomer has the area between two lone pairs and a hydrogen (HLL) pointed towards the first molecule. 51

4.8 (a) GGA exchange enhancement factor $F_x(s)$, optimized without constraints to reproduce energies calculated with the MB-pol potential [6] of 300 water monomers (CUSTOM-1B (unconstr.), red) and that of vdW-DF-cx (LV-PW86R, blue) and (b) energies of 300 monomers calculated with the CUSTOM-1B functional (red) and with the vdW-DF1-cx functional (blue) against the MB-pol reference energies. Dots indicate parameter values, and lines represent the cubic interpolations between the points. The RMS errors are shown in the legends. . . . 53
4.9 The electron density $\rho(k_F, k_G)$ (Eq. 4.20) for a monomer at its equilibrium geometry, with a 300 Ry mesh cutoff (a) and a 600 Ry mesh cutoff (b). 55
4.10 Changes in the electron density $\rho(k_F, k_G)$ upon moving from the monomer equilibrium geometry of vdw-DF-cx functional to that of MB-pol, with a 300 Ry mesh cutoff (a) and a 600 Ry mesh cutoff (b). The lines represent constant values of s. The equilibrium geometries can be found in Table 4.2. 56
4.11 Changes in the electron density $\rho(s)$ (Eq. 4.21) upon changing from the water monomer equilibrium geometry of vdw-DF-cx functional to that of MB-pol, with (a) 300 Ry and (b) 600 Ry mesh cutoffs. Also represented are the changes in the electron density $\rho(s)$ multiplied by the LDA exchange energy density ε_x^{LDA} to show the weight of the changes. 57
4.12 61
4.12 (a) Plot of all forms of $F_x(s)$ used to calculate the average $F_x^0(s)$ (Eq. 4.12b) and the covariance matrix $Cov(s, s')$ (Eq. 4.24) in all Bayesian constrained DPPS optimizations presented in this thesis and (b) plot of $F_x^0(s) \pm$ the diagonal of $Cov(s, s')$, which is plotted in (c). The contour values in (c) are at increments of 1.35×10^{-3}. 62
4.13 The plots in (a) and (b) are the same as those in Figures 4.13a and 4.8b, with the exception that Bayesian constraints were used to fit the functional CUSTOM-1B. 63

4.14 (a) GGA exchange enhancement factor $F_x(s)$, optimized without constraints to reproduce 2-body energies calculated with the MB-pol potential [6] of 330 dimers (CUSTOM-2B (unconstr.), red) and that of vdW-DF-cx (LV-PW86R, blue) and (b) 2-body energies of the same 330 dimers calculated with the CUSTOM-2B functional (red) and with the vdW-DF1-cx functional (blue) plotte against the MB-pol reference energies. The average errors $\pm$ RMS errors are shown in the legends. 64
4.15 Changes in the electron density $\rho(k_F, k_G)$ (a) and $\rho(s)$ (b) upon the formation of a dimer at the equilibrium geometry, calculated with a 600 Ry mesh cutoff. The lines represent constant values of s. 66
4.16 The plots in (a) and (b) are the same as those in Figures 4.14a and 4.14b, with the exception that Bayesian constraints were used to fit the functional CUSTOM-2B. 67
4.17 (a) GGA exchange enhancement factor $F_x(s)$, optimized without constraints to reproduce MB-pol [6] 3-body energiesl of 100 water trimers (CUSTOM-3B (unconstr.), red) and that of vdW-DF-cx (LV-PW86R, blue) and (b) 3-body energies of 100 trimers calculated with the CUSTOM-1B functional (red) and with the vdW-DF1-cx functional (blue) plotted against the MB-pol reference energies. The average errors $\pm$ RMS errors are shown in the legends. 69
4.18 The plots in (a) and (b) are the same as those in Figures 4.17a and 4.17b, with the exception that Bayesian constraints were used to fit the functional CUSTOM-3B. 70
4.19 71
4.19 (a) GGA exchange enhancement factor $F_x(s)$, optimized without constraints to reproduce 2-body and 3-body energies calculated with the MB-pol potential [6] of 330 water dimers and 100 water trimers, respectively (CUSTOM-MB (unconstr.), red) and that of vdW-DF-cx (LV-PW86R, blue), (b) 2-body energies of 330 dimers and (c) 3-body energies of 100 trimers calculated with the CUSTOM-1B functional (red) and with the vdW-DF1-cx functional (blue) against the MB-pol reference energies. The average errors $\pm$ RMS errors are shown in the legends. 72
4.20 74

4.20 The plots in (a), (b), and (c) are the same as those in Figures 4.19a,4.19c and 4.17b, respectively with the exception that Bayesian constraints were used to fit the functional CUSTOM-MB and (d) 100 monomer energies calculated with the CUSTOM-MB functional (red) and with the vdW-DF1-cx functional (blue) plotted against the MB-pol reference energies. The average errors $\pm$ RMS errors are shown in the legends of (a)-(c), and the RMS errors are shown in the legend of (d). 75

5.1 Dependence of (a) the average simulation pressure $\langle P \rangle$, (b) the average order parameter $\langle \zeta \rangle$, and the position of (c) the first maximum $g_{OO}^{1^{st}max}$ and (d) the first minimum $g_{OO}^{1^{st}min}$ in the oxygen-oxygen RDF on average simulation temperature $\langle T \rangle$. All simulations were done with the vdW-DF-cx functional, 32 water molecules per simulation cell, the $^{(p)}$dζ + p the basis set, a 200 Ry mesh cutoff, and a 5×10^{-4} Bohr^{-3} DM tolerance. The error bars were calculated using the first term of Eq. 5.2, and the slopes of the linear fits are given in Table 5.1. 83

5.2 Comparison of average pressures $\langle P \rangle$ over 20,000 step AIMD trajectories, with different functionals (red), and their TI estimates (blue, Eq. 5.12), using a single set of 200 snapshots produced by an AIMD simulation with the vdW-DF-cx functional, with 32 molecules per unit cell, the $^{(p)}$dp+p basis set, a 300 Ry mesh cutoff, and a 5×10^{-4} Bohr^{-3} DM tolerance. The AIMD value for vdW-DF-cx was averaged over the full trajectory of this simulation. The AIMD values and error bars for PBE and vdW-DF^{PBE} were taken from Ref. [20], which used 64 molecules per unit cell, and the values for vdW-DF and VV were taken from Ref. [129], which used 200 molecules per unit cell. In both references, the $^{(p)}$dζ+p basis set was used. Error bars were not provided in Ref. [129]. 90

5.3 92

5.4 The difference between the $g_{OO}(r)$ calculated from an AIMD trajectory generated at 300 K with the $^{(s)}$qζ+dp basis and the one calculated with the $^{(p)}$dζ+p basis plotted along with the difference estimated with the TI method, using 200 snapshots from the $^{(p)}$dζ+p simulation. All calculations used the vdW-DF-cx functional, a 200 Ry mesh cutoff, 32 molecules per unit cell, and a 5×10^{-4} Bohr^{-3} DM tolerance. 92

5.5 Plots of $\langle P \rangle$ (a), $\langle \zeta \rangle$ (b), and the first maximum (c) and first minimum position (d) of $g_{OO}(r)$ for different system sizes: 32, 64, and 128 molecules per unit cell. All simulations are done with a $^{(p)}$dζ+p basis set, a 200 Ry mesh cutoff, and a DM tolerance of 1×10^{-4} Bohr^{-3}. Error bars are determined by Eq. 5.2 95

5.6 Plots of $\langle P \rangle$ (a), $\langle \zeta \rangle$ (b), and the first maximum (c) and first minimum position (d) of $g_{OO}(r)$ for four different basis sets: $^{(p)}$dζ+p, tζ+dp, $^{(s)}$qζ+dp and qζ+dp. All simulations are done with 32 water molecules per unit cell, a 200 Ry mesh cutoff, and a DM tolerance of 1×10^{-4} Bohr^{-3}. 96

5.7 Plots of $\langle P \rangle$ (a), $\langle \zeta \rangle$ (b), and the first maximum (c) and first minimum position (d) of $g_{OO}(r)$ for three different mesh cutoffs: 200 Ry, 300 Ry, and 400 Ry. All simulations are done with 32 water molecules per unit cell, a $^{(p)}$dζ+p basis set, and a DM tolerance of 1×10^{-4} Bohr^{-3}. 97

5.8 Plots of $\langle P \rangle$ (a), $\langle \zeta \rangle$ (b), and the first maximum (c) and first minimum position (d) of $g_{OO}(r)$ for different DM tolerances (2×10^{-4}, 1×10^{-4}, 0.5×10^{-4}, 0.2×10^{-4}, and 0.1×10^{-4} Bohr^{-3}). All simulations are done with 32 water molecules per unit cell, a 200 Ry mesh cutoff, and a $^{(p)}$dζ+p basis set. 98

6.1 The five isomers of the H_2O hexamer for which we analyze relative binding energies with different DFT functionals: (a) prism; (b) cage; (c) book; and (d) bag: and (e) cyclic. Red and gray spheres represent O and H atoms, with connecting lines showing hydrogen bonds. . . 102

6.2 Binding energies (in kcal/mol) of the prism, cage, book, bag, and cyclic water hexamer isomers shown in Fig. 6.1, relative to the binding energy of the prism isomer. The energies calculated for MP2 geometries are plotted, using MB-POL and three different DFT functionals: vdW-DF-cx, CUSTOM-2B-Δ_1, and CUSTOM-MB-Δ_1. 103
6.3 MAD with respect to MB-pol of the total energy per molecule of 200 different 6-molecule water clusters taken from liquid water snapshots. The MAD was calculated with the following functionals: vdW-DF-cx (BH), vdW-DF-cx-Δ_1 (BH-Δ_1), CUSTOM-2B (C-2B), CUSTOM-2B-Δ_1 (C-2B-Δ_1), CUSTOM-MB (C-MB) and CUSTOM-MB-Δ_1 (C-MB-Δ_1). The liquid water snapshots were taken from a 15 ps AIMD simulation with the vdW-DF-cx functional at 300 K. 106
6.4 This figure is taken from Ref. [83]. It plots the MAD with respect to QMC of 50 water snapshots from a vdW-DF2 [62] Path Integral Molecular Dynamics (PIMD) simulation, done at 300 K with 64 molecules per unit cell. Energies are calculated with 4 different DFT functionals with 1-body (red), 2-body (green), or 3-body (purple) energies corrected. Uncorrected energies are labelled Bare (blue). . . 107
6.5 110
6.5 Oxygen-oxygen RDF $g_{OO}(r)$ (a), oxygen-hydrogen RDF $g_{OH}(r)$ (b), and hydrogen-hydrogen RDF $g_{HH}(r)$ (c) resulting from DFT AIMD simulations with the following functionals: vdW-DF-cx (blue), vdW-DF-cx-Δ_1 (blue dash), CUSTOM-2B (magenta), CUSTOM-2B-Δ_1 (magenta dash), CUSTOM-MB (green), CUSTOM-MB-Δ_1 (green dash). For all DFT simulations, 128 molecules per unit cell, a $^{(p)}$dζ+p basis set, 300 Ry mesh cutoff, and 0.5×10^{-5} Bohr^{-1} DM tolerance were used. Also plotted for reference are the RDFs obtained from an MB-POL MD simulation (red) in Figures (a)-(c) and the experimental RDF $g_{OO}(r)$ (black) in (a) and for comparison. Data for the experimental RDFs was taken from Ref. [117]. 111
6.6 113

6.6 Interaction energy curves of two rigid water monomers as a function of the O-O distance (r) for an H-bonded-dominant configuration (6.6a) and a vdW-dominant configuration (6.6b). The dimer is not allowed to relax, and the only degree of freedom is r. In both graphs the energies have been shifted to have the zero at the largest separation. Orientations of the monomers are shown for each configuration. All energies were calculated with the $^{(p)}$dζ+p basis set. 114

6.7 Taken from Ref. [129]. Plot of the O-O RDF $g_{OO}(r)$ obtained with vdW-DF1, referred to as DRSLL here (black line), to show the collapse of the second coordination shell with this functional. This was taken from a simulation of water with a density of $1.0\ g/cm^3$ at 300 K, with a $^{(p)}$dζ+p basis set and 64 molecules per unit cell. The vdW-DF1 (DRSLL) result is compared to that of vdW-DF1 with PBE exchange (DRSLL-PBE, red), and the results from the experimental data of Refs. [45] (dotted line) and [119] (dashed line). 115

6.8 Decomposition of g_{OO}(r) from Fig. 6.5a. Top: first H-bonded neighbors, constituting the first coordination shell. Center: second H-bonded neighbors, constituting the second coordination shell. Bottom: remaining molecules which include vdW-bonded neighbors in the interstitial region and H-bonded molecules beyond the second coordination shell. 116

6.9 TI estimation of oxygen-oxygen RDF $g_{OO}(r)$ with CUSTOM-MB-Δ_1 and the $^{(s)}$qζ+dp basis set (black), using 200 snapshots of the AIMD trajectory generated with CUSTOM-MB-Δ_1 and the $^{(p)}$dζ+p basis set, compared with the AIMD/ $^{(p)}$dζ+p (green, dot-dash) and MB-pol (red) results. All parameters other than the basis set are the same as those described in Fig. 6.5. 118

6.10 119

6.10 Histograms of the intramolecular distance between an oxygen atom and a hydrogen atom r_{OH} (6.10a) and the H-O-H angle for each molecule θ_{HOH} (6.10b). Averages are provided in the legend and are taken over 300,000 values. These values are for 10 molecules from each step of the 30,000 trajectories from the AIMD simulation with the CUSTOM-2B functional and the monomer corrected version CUSTOM-2B-Δ_1. 120

List of tables

4.1 Mean errors of 300 1-body energies, 330 2-body energies, and 100 3-body energies calculated with the vdW-DF-cx functional and all CUSTOM functionals described here. The monomers, dimers, and trimers are those of the sets used in the optimizations. Each entry gives the mean deviation of the energy from the reference energy $\pm$ the RMS fluctuation of the energy deviation per atom. In the case of the 1-body terms, only the RMS values are included, since the average energy error for each functional is $\equiv 0$. All values are in meV/atom. 77

4.2 Equilibrium distances between O and H r_{OH}, equilibrium angle between two O-H bonds, and electric dipole moment at the reference equilibrium geometry calculated with MB-pol of a water monomer, calculated with the vdW-DF-cx functional and CUSTOM functionals optimized with Bayesian constrained DPPS. To the right of each calculated values, the percent errors of values taken with respect to the values of MB-pol are displayed. In all calculations, a 300 Ry mesh cutoff, 20 Å lattice constant, and $^{(s)}$qζ-dp basis set were used. . . . 78

5.1 Values from AIMD liquid water simulations of the average simulation pressure $\langle P \rangle$, the average order parameter $\langle \zeta \rangle$, and the position of the first maximum $g_{OO}^{1^{st}max}$ and the first minimum $g_{OO}^{1^{st}min}$ in the oxygen-oxygen RDF $g_{OO}(r)$. All simulations were done with the vdW-DF-cx functional, 32 water molecules per unit cell, the $^{(p)}$dζ + p basis set, a 300 Ry mesh cutoff, and a 5×10^{-4} Bohr^{-3} DM tolerance. These numbers are used to calculate the linear fit of each property plotted against the average simulation temperature $\langle T \rangle$ (Fig. 5.1). The slopes $d\alpha/d\langle T \rangle$ of the fits are included in the bottom table. 84

5.2 Average temperature $\langle T \rangle$ of AIMD simulations performed for convergence tests, along with temperature adjusted (Eq. 5.1) values of the following properties: average simulation pressure $\langle P \rangle$, average order parameter $\langle \zeta \rangle$ and the positions of the first maximum and minimum in the oxygen-oxygen RDF $g_{OO}(r)$. All simulations were done with the vdW-DF-cx functional, and unless otherwise noted, 32 molecules per unit cell, the $^{(p)}$dζ+p basis set, a 200 Ry mesh cutoff, and a 5×10^{-4} Bohr^{-3} DM tolerance. All values in the tests for number of molecules, basis set, and mesh cutoff are adjusted for T=311 K, and in the tests for DM tolerance they are adjusted for T=323 K. All temperatures are in K, pressures are in kBar, and DM tolerances are in Bohr^{-3}. 85

5.3 Positions $r_{OO}^{1^{st}max}$ of the first maximum and r_{OO}^{min} of the first minimum, and the heights $g_{OO}^{1^{st}max}$ of the first peak and g_{OO}^{min} of the first valley of $g_{OO}(r)$, obtained with the following methods: (i) an AIMD simulation using the $^{(p)}$dζ+p basis set, (ii) an AIMD simulation using the $^{(s)}$qζ+dp basis set, and (iii) the TI estimation of the $^{(s)}$qζ+dp RDF, using 200 snapshots from the $^{(p)}$dζ+p simulation. All calculations were performed with the vdW-DF-cx functional, a 200 Ry mesh cutoff, 32 molecules per unit cell, and a 5×10^{-4} Bohr^{-3} DM tolerance. The target temperature was 300 K for both AIMD simulations. All distances are given in Å. 93

6.1 Average values of the following properties obtained from AIMD simulations with different functionals: average pressure $\langle P \rangle$, average temperature $\langle T \rangle$, and average order parameter $\langle \zeta \rangle$. The values are averaged over the last 10 ps of each simulation. For reference, the value of $\langle \zeta \rangle$ obtained with the MB-pol simulation is 0.0210 nm. A $^{(s)}$dζ+p basis set, 300 Ry mesh cutoff, and a 0.5×10^{-4} Bohr^{-3} DM tolerance were used for all AIMD simulations. 108

6.2 RDF properties obtained from the same AIMD simulations described in Table 6.1: positions $r_{OO}^{1^{st}max}$ and $r_{OO}^{2^{nd}max}$ of the first and second maximum, position r_{OO}^{min} of the first minimum, the heights $g_{OO}^{1^{st}max}$ and $g_{OO}^{2^{nd}max}$ of the first and second peaks, and the height g_{OO}^{min} of the first valley of $g_{OO}(r)$ (Fig. 6.5a). The RDFs are obtained from the trajectories over 15 ps of simulation. The last row shows the TI estimate of $g_{OO}(r)$ with CUSTOM-MB-Δ_1 and a $^{(}s)$qζ+dp basis set, using 200 snapshots from the AIMD trajectory using CUSTOM-MB-Δ_1 and a $^{(}p)$dζ+p basis set. 109

Chapter 1

Introduction

While computational methods of studying molecular systems have continued to advance over the past years, obtaining a good balance between computational efficiency and accuracy in estimations of electronic interactions remains a challenge. As methods improve in their ability to accurately represent electronic interactions, their rise in computational cost typically limits their use to small systems. Consequently, chemical accuracy[1] remains out of reach for condensed systems.

The balance between accuracy and efficiency can be improved for specific systems through the parameterization of efficient force fields and models derived from first principles (also referred to as *ab initio* models). Once a dataset of properties has been calculated using a high precision method for small systems, more efficient models can be optimized to the dataset by adjusting a set of parameters. Unfortunately, the process of parameterization of a complex model is complicated and involves choosing largely arbitrary functional forms that depend on many parameters, followed by a lengthy and difficult trial and error optimization process. The balance between the number of parameters and the size of the fitted data sets involves decisions that are difficult and subjective, yet critical for obtaining the best result. In this thesis, a general and powerful optimization scheme that provides an efficient way to obtain high-quality parameters, termed "Data Projection onto Parameter Space" (DPPS), is presented.

We have focused on the application of DPPS to the optimization of models within the framework of Density Functional Theory (DFT) for water systems. Water, arguably the most important liquid on earth, is a surprisingly complex substance, in-

[1] Chemical accuracy is usually defined as within 1kcal/mol of the experimental value.

volving a delicate interplay of covalent, electrostatic, and dispersion interactions. A proper description of liquid water is very sensitive to the relative strengths of these forces. Thus, it is not surprising that DFT, a universal and non-empirical method, has had a particularly hard time in describing all these interactions with the required accuracy. In fact, as of today, DFT simulations are unable to match the success of empirical force fields in simulating a wide range of effects and anomalies seen in water. However, empirical methods are not necessarily reliable outside the range for which they have been fitted, such as in deeply under-cooled water, a state which has been predicted to have a liquid-liquid transition with a behavior that is connected to anomalies at higher temperatures [112]. Therefore, an accurate representation with DFT remains a very important challenge in the study of the intricate structure and properties of liquid water.

The work presented in this thesis involves uncovering in detail the deficiencies of present density functionals in describing water systems and the development of an optimized exchange functional within the Generalized Gradient Approximation (GGA). The structure of liquid water has been widely investigated in past years with GGA-DFT functionals [33, 58, 84, 116, 46]. One of the problems with GGA-DFT is that it ignores non-local electron correlation effects that are responsible for van der Waals (vdW) interactions, commonly referred to as dispersion interactions, which are thought to be crucial in some systems, including different phases of water. The inclusion of a vdW functional is essential for an accurate description of long-range interactions in water, and is therefore important for correctly predicting the structure of liquid water as well as the energetics of water clusters and ice. In addition, it has been shown that the proper choice of the exchange functional in combination with a vdW functional is crucial for a proper description of water [129, 76]. Therefore, by including a vdW functional along with an optimized version of GGA exchange, an improved description of water with DFT is expected.

When considering the optimization of a functional, it is necessary to have a reliable and thorough reference dataset of properties from a highly accurate method. To obtain an appropriate set of interaction energies for water systems, we used the MB-pol potential [6], which is the most accurate potential for water available today and is the only many-body potential that accurately predicts the properties of water ranging from the gas to the condensed phase.

MB-pol takes advantage of the fact that the many-body expansion of the interaction energy [75] for water converges rapidly [133, 133, 43, 46, 134, 22, 41, 35, 48].

It is well known that the potential energy of a system with N interacting molecules can be expressed in terms of the many-body expansion:

$$V_N(x_1,...,x_N) = \sum_a V^{(1B)}(x_a) + \sum_{a>b} V^{(2B)}(x_a,x_b) + \sum_{a>b>c} V^{(3B)}(x_a,x_b,x_c) + ... + V^{(NB)}(x_1,...,x_N). \quad (1.1)$$

Here x_a denotes the coordinates of all the atoms of the a^{th} molecule and $V^{(nB)}$ are the n-body interactions with $1 \leq n \leq N$, which are defined recursively as

$$V^{(nB)}(x_1,...,x_n) = V_n(x_1,...,x_n) - \sum_a V^{(1B)}(x_a) - \sum_{a>b} V^{(2B)}(x_a,x_b) - ... - \sum_{a_1>a_2>...>a_{n-1}} V^{((n-1)B)}(x_{a_1},x_{a_2},...,x_{a_{n-1}}). \quad (1.2)$$

The rapid convergence of Eq 1.2 allows for the potential energy surface to be estimated as a sum of the low-order terms. Studies on theoretical modelling of water have shown that terms beyond the 3-body do not make a significant contribution to the total potential energy [78]. Thus, a high-quality potential energy surface for water could be developed by optimizing up to the 3-body term to values obtained with a highly accurate electronic structure method. The development of the MB-pol potential is based on this strategy, and includes 2-body and 3-body terms optimized for thousands of dimers and trimers, respectively, to the values obtained with the coupled cluster singles and doubles (triplets) or CCSD(T) method, currently considered the gold standard in quantum chemistry. For calculations of the 1-body terms, MB-pol incorporates the energy surface by Partridge and Schwenke [94].

For the work presented in this thesis, using the many-body representation of the interaction energy, we have optimized a vdW density functional (vdW-DF) to the first three energy terms of Eq. 1.1 for water. The functional was optimized with DPPS, and only the GGA exchange part of the functional was parameterized. The DPPS-optimized functionals were then applied to gas-phase water cluster calculations and *ab initio* molecular dynamics (AIMD) simulations of liquid water to test the importance of the corrections to each of the first three terms in the many-body expansion in accurately representing these systems.

In Chapter 2, the basic concepts involved in DFT and the functionals discussed in this thesis, as well as the SIESTA method [118], which was the main tool for the calculations that we have done, are introduced.

Chapter 3 discusses the basic properties of liquid water and important factors to take into account when pursuing a DFT functional that will accurately reproduce its experimental structure.

Chapter 4 introduces the methodology of DPPS for parameterization, focusing on the optimization of a GGA exchange functional. The process of optimization of up to 3-body terms for water are presented and discussed.

In Chapter 5, the results of several AIMD simulations of liquid water are presented and analyzed to determine the convergence of simulation parameters for later calculations. A new method for convergence testing based on thermodynamic integration (TI) is presented, and results from this method are compared with the AIMD results.

In Chapter 6, new functionals resulting from the optimization of the GGA exchange, following the methodology discussed in Chapter 4, are tested on gas phase water clusters and AIMD simulations of liquid water. Parameters that were determined in Chapter 5 through convergence testing were used in the calculations discussed in this chapter.

In Chapter 7, the main conclusions about the results presented in this thesis and ideas for future extensions of the work are discussed.

Chapter 1

Introducción

Se puede establecer un equilibrio entre precisión y coste computacional para sistemas específicos a través de la parametrización de campos de fuerzas eficientes y modelos derivados de primeros principios (también llamados modelos de ab initio). Una vez que un conjunto de datos de un método de alta precisión está obtenido para sistemas pequeños, se pueden optimizar modelos más eficientes ajustando unos parámetros. Desafortunadamente, el proceso de parametrización de un modelo complejo es complicado y depende de elecciones arbitrarias de formas funcionales que dependen de muchos parámetros, seguido por un proceso largo y difícil de optimización. Para mantener el equilibrio entre el número de parámetros y el tamaño del conjunto de datos se requieren decisiones que son difíciles y subjetivas, pero críticas para obtener el mejor resultado. En esta tesis, se presenta un esquema general y poderoso que proporciona una manera eficiente de obtener parámetros de alta calidad, llamado en inglés “Data Projection onto Parameter Space” o “Proyección de Datos al Espacio de Parámetros” en español (DPPS).

Nos Hemos enfocado en la aplicación de DPPS a la optimización de modelos dentro de la estructuras de la teoría de funcionales de densidad (DFT) para sistemas de agua. El Agua, que uno podría decir que es el líquido más importante de la tierra, es una sustancia asombrosamente compleja, que depende de una combinación delicada de interacciones covalentes, electrostáticas, y dispersivas. Una descripción apropiada del agua líquida es muy sensible a las magnitudes relativas de estas interacciones. Así, no es sorprendente que DFT, un método universal y no empírico, haya tenido problemas para describir estas interacciones con la precisión requerida. De hecho, hoy en día las simulaciones con DFT no pueden competir con el éxito que han tenido los campos de fuerzas empíricos para simular un rango

grande de efectos y anomalías vistas en el agua. Sin embargo, los métodos empíricos no son necesariamente fiables fuera del rango para que han sido ajustados, por ejemplo en agua subenfriada (?), un estado en el que está previsto tener una transición liquido-liquido con un comportamiento que está conectado a las anomalías vistas a temperaturas más altas [112]. Por lo tanto, una representación con DFT sigue siendo un desafío importante en el estudio de la estructura intricada y las propiedades de agua.

El trabajo presentado en esta tesis involucra destapar en detalle las deficiencias de funcionales de densidad para describir sistemas de agua y el desarrollo de un funcional de intercambio optimizado dentro de la "Generalized Gradient Approximation" (GGA). La estructura del agua líquida ha sido investigada extensamente en los últimos años con funcionales GGA DFT [33, 58, 84, 116, 46]. Uno de los problemas con GGA-DFT es que ignora los efectos no-locales de correlación de los electrones que son responsables de interacciones de van der Waals (vdW), que son importantes en unos sistemas, incluyendo fases diferentes de agua. La inclusión de un funcional de vdW es esencial para obtener una descripción precisa de las interacciones a largas distancias en agua, y por lo tanto es importante para predecir correctamente la estructura del agua líquida tanto como las energías de clústeres de agua y hielo. Además, otros estudios han enseñado que la elección apropiada de un funcional de intercambio en combinación con un funcional de vdW es crucial para obtener una descripción correcta del agua [129, 76]. Por lo tanto, al incluir un funcional de vdW junto con una versión optimizada de intercambio GGA, se espera una mejor descripción del agua con DFT.

Cuando uno está considerando la optimización de un funcional, es necesario tener un conjunto de datos que sea fiable y completo. Para obtener un conjunto de datos apropiado de energía de interacción del agua, usamos el potencial MB-pol [6], considerado el potencial más preciso que existe para agua y es el único potencial que predice de manera precisa las propiedades de agua, incluyendo las fases, desde las de gas hasta las condensadas.

El potencial MB-pol aprovecha el hecho de que la expansión de muchos cuerpos de energía de interacción para agua [75] converge rápido [133, 133, 43, 46, 134, 22, 41, 35, 48]. Es bien sabido que el potencial de energía de un sistema con N moléculas interactuando se puede expresar en términos de la expansión de

muchos cuerpos:

$$V_N(x_1,...,x_N) = \sum_a V^{(1B)}(x_a) + \sum_{a>b} V^{(2B)}(x_a,x_b) + \sum_{a>b>c} V^{(3B)}(x_a,x_b,x_c) + ... + V^{(NB)}(x_1,...,x_N). \quad (1.1)$$

Aquí x_a indica las coordenadas de todos los átomos de la a° molécula y $V^{(nB)}$ son los n-cuerpo interacciones con $1 \leq n \leq N$, que se definen recursivamente como

$$V^{(nB)}(x_1,...,x_n) = V_n(x_1,...,x_n) - \sum_a V^{(1B)}(x_a) - \sum_{a>b} V^{(2B)}(x_a,x_b) - ... - \sum_{a_1>a_2>...>a_{n-1}} V^{((n-1)B)}(x_{a_1},x_{a_2},...,x_{a_{n-1}}). \quad (1.2)$$

La convergencia rápida de Eq 1.2 permite una representación de una superficie de energía compuesta de la suma de solo los primeros términos. Estudios de modelos teóricos de agua han enseñado que los términos después los de 3-cuerpos no hacen una contribución significante a la energía total [78]. Entonces, se podría desarrollar un potencial para el agua de alta calidad optimizando los primeros 3 términos a valores obtenidos con un método de alta precisión. El desarrollo de MB-pol se basa en esta estrategia, e incluye términos de 2-cuerpos y de 3-cuerpos optimizados a miles de dímeros y trímeros, respectivamente, a los valores obtenidos con el método de CCSD(T), lo que se considera hoy en día la mejor precisión en química cuántica. Para los cálculos de los términos de 1-cuerpo, MB-pol incorpora la superficie de Partridge y Schwenke [94].

En el trabajo presentado en esta tesis, usando la representación de muchos cuerpos de energía de interacción, hemos optimizado un funcional de densidad que incluye vdW (vdW-DF) a los primeros tres términos de Eq. 1.1 para agua. El funcional fue optimizado con DPPS y solamente la parte de intercambio GGA fue parametrizada. Aplicamos los funcionales optimizados con DPPS a cálculos con clústeres de agua en la fase de gas y a dinámica molécular ab initio (AIMD) de agua líquida para comprobar la importancia de las correcciones a cada uno de los primeros tres términos de la expansión de muchos cuerpos, precisamente representando estos sistemas.

Chapter 2

Theory

This chapter introduces the theoretical background involved in the work presented in this thesis, focusing on the path to exchange and correlation functionals in Density Functional Theory (DFT), mostly following Ref. [72]. DFT has been the most popular method used for electronic structure calculations in solid-state physics for many years. By using functionals that depend on electron density and mapping an interacting system to a non-interacting one, DFT provides a good balance between accuracy and computational cost. A significant part of the work presented in following chapters involves the optimization of exchange density functionals, which are used to estimate exchange interactions in DFT. The optimizations and applications of these density functionals were carried out with the SIESTA [118] implementation of DFT, which is introduced in Section 2.6.

2.1 The Hartree-Fock approximation

The Hartree-Fock (HF) approximation is a standard approach to solving the Schrödinger equation in order to find the ground state energy of a system of electrons and nuclei. In the non-relativistic case, using the Born-Oppenheimer approximation [16], the time-independent Schrödinger equation is

$$\hat{H}\Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) = E\Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N)\,. \tag{2.1}$$

Here the values of $\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N$ are the positions of N electrons. The electron Hamiltonian operator in atomic Hartree units is [73]

$$\hat{H} = -\frac{1}{2}\sum_i \nabla_i^2 + \sum_i \sum_I \frac{Z_I}{\mid \mathbf{r}_i - \mathbf{R}_I \mid} + \frac{1}{2}\sum_{i \neq j} \frac{1}{\mid \mathbf{r}_i - \mathbf{r}_j \mid} + \frac{1}{2}\sum_{I \neq J} \frac{Z_I Z_J}{\mid \mathbf{R}_I - \mathbf{R}_J \mid}$$
$$= \hat{T} + \hat{V}_{ext} + \hat{V}_{int} + \hat{V}_{II}. \qquad (2.2)$$

Here the first term $\hat{T}$ is the kinetic energy of the electrons. The second term $\hat{V}_{ext}$ is the interaction energy of the electrons with the atomic nuclei and with any other external field. The third term $\hat{V}_{int}$ is the electron-electron interaction, and the last term $\hat{V}_{II}$ is the classical interaction between atomic nuclei. Electrons, with positions $\mathbf{r}_i$, are denoted by lower case subscripts and nuclei, with charges Z_I and positions $\mathbf{R}_I$, are denoted by upper case subscripts.

In the HF method, the total wave function is approximated in terms of N single-electron wave functions. Because we are dealing with fermions, this wave function should be anti-symmetric with respect to an interchange of any set of space-spin coordinates. In order to achieve anti-symmetry, the HF method defines the wave function as a Slater determinant, i.e.,

$$\Phi = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\mathbf{r}_1, \sigma_1) & \phi_1(\mathbf{r}_2, \sigma_2) & \phi_1(\mathbf{r}_3, \sigma_3) & ... \\ \phi_2(\mathbf{r_1}, \sigma_1) & \phi_2(\mathbf{r_2}, \sigma_2) & \phi_2(\mathbf{r_3}, \sigma_3) & ... \\ \phi_3(\mathbf{r_1}, \sigma_1) & \phi_3(\mathbf{r_2}, \sigma_2) & \phi_3(\mathbf{r_3}, \sigma_3) & ... \\ . & . & . & ... \end{vmatrix}, \qquad (2.3)$$

where $\phi_i(\mathbf{r}_j, \sigma_j)$ are single particle "spin-orbitals". They are products of functions of position $\psi_i^{\sigma}(\mathbf{r}_j)$ and functions of spin variables $\alpha_i(\sigma_j)$.

The total energy of the system can then be found by calculating the expectation value of the Hamiltonian, so that the HF energy is

$$E_{HF} = \langle \Phi | \hat{H} | \Phi \rangle = \sum_{i,\sigma} \int d\mathbf{r} \psi_i^{\sigma *}(\mathbf{r}) \left[-\frac{1}{2}\nabla^2 + V_{ext}(\mathbf{r}) \right] \psi_i^{\sigma}(\mathbf{r}) + E_{II}$$
$$+ \frac{1}{2} \sum_{i,j,\sigma_i,\sigma_j} \int d\mathbf{r} d\mathbf{r}' \psi_i^{\sigma_i *}(\mathbf{r}) \psi_j^{\sigma_j *}(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} \psi_i^{\sigma_i}(\mathbf{r}) \psi_j^{\sigma_j}(\mathbf{r}')$$
$$- \frac{1}{2} \sum_{i,j,\sigma} \int d\mathbf{r} d\mathbf{r}' \psi_i^{\sigma *}(\mathbf{r}) \psi_j^{\sigma *}(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} \psi_j^{\sigma}(\mathbf{r}) \psi_i^{\sigma}(\mathbf{r}'). \qquad (2.4)$$

In the first term, the single-body terms are grouped together, and the second term is the electrostatic nucleus-nucleus interaction. The third term is the direct interaction, which is the classical Coulomb energy written in terms of orbitals. This term involves the direct interaction of electron i at position $\mathbf{r}$ with the electron j at position $\mathbf{r}'$. The last term is the exchange interaction, in which electrons i and j are interacting with each other at interchanged positions.

The variational theorem [32] shows that the ground state Φ_0 has the lowest energy E_0. Then any state Φ must result in an energy greater than or equal to the ground state energy, i.e.,

$$E[\Phi] \geq E_0. \tag{2.5}$$

Thus, the ground state Φ_0 can be determined by minimization of the total energy with respect to all the degrees of freedom of the wave function Φ. Doing so leads to the Hartree-Fock equations, which are also known as the self-consistent field (SCF) equations:

$$\left[-\frac{1}{2}\nabla^2 + V_{ext}(\mathbf{r}) + \underbrace{\sum_{j,\sigma_j} \int d\mathbf{r}' \psi_j^{\sigma_j *}(\mathbf{r}') \psi_j^{\sigma_j}(\mathbf{r}') \frac{1}{|\mathbf{r}-\mathbf{r}'|}}_{\hat{V}_{Hartree}} \right.$$
$$\left. \underbrace{-\sum_j \int d\mathbf{r}' \psi_j^{\sigma *}(\mathbf{r}') \psi_i^{\sigma}(\mathbf{r}') \frac{1}{|\mathbf{r}-\mathbf{r}'|} \frac{\psi_j^{\sigma}(\mathbf{r})}{\psi_i^{\sigma}(\mathbf{r})}}_{\hat{V}_x^{i,\sigma}} \right] \psi_i^{\sigma}(\mathbf{r}) = \varepsilon_i^{\sigma} \psi_i^{\sigma}(\mathbf{r}) \tag{2.6}$$

These equations describe non-interacting electrons under the influence of a mean field potential, which includes the Coulomb potential $\hat{V}_{Hartree}$ and the non-local exchange potential $\hat{V}_x^{i,\sigma}$. Improvements of the wave function and energy resulting from this method can be achieved by introducing extra degrees of freedom in the wave function, which always lowers the energy for any state, according to a theorem by MacDonald [71]. The lowering of the energy with respect to the Hatree-Fock energy is called the correlation energy E_c. Taking E_c into account with this method, however, becomes too computationally expensive for practical uses as it scales dramatically with the number of electrons considered. The computational cost can be greatly reduced by using DFT to find an approximate solution to the Schrödinger equation, as outlined in the following sections.

2.2 Density functionals

In the solution of the Schrödinger equation using the Hartree-Fock method, knowledge of a $3N$ dimensional wave function is needed. However, with DFT, an estimation of the energy and other properties of a system is made using the ground state density of particles $n_0(\mathbf{r})$, a scalar function of only one coordinate, and the use of the many-body wave function becomes unnecessary. This section gives a brief overview of the basic concepts behind this method.

Hohenberg and Kohn

DFT is based on theorems originally proven by Hohenberg and Kohn (HK) [44], which treat the energy as a variational functional of the electron density $n(\mathbf{r})$. The HK theorems can be stated as follows:

Theorem I: If two systems of interacting particles are trapped in potentials $V_{ext}^{(1)}(\mathbf{r})$ and $V_{ext}^{(2)}(\mathbf{r})$ and both systems have the same ground state density $n_0(\mathbf{r})$, then the two potentials are equal, except for a constant, i.e., $V_{ext}^{(1)}(\mathbf{r}) - V_{ext}^{(2)}(\mathbf{r}) = const$.

Proof: Let us suppose that $V_{ext}^{(1)}(\mathbf{r})$ and $V_{ext}^{(2)}(\mathbf{r})$ differ by more than a constant and lead to the same ground state density $n_0(\mathbf{r})$. There would then be two different Hamiltonians, $\hat{H}^{(1)}$ and $\hat{H}^{(2)}$ and two different ground state wave functions, $\Psi^{(1)}$ and $\Psi^{(2)}$, which give us the same ground state density $n_0(\mathbf{r})$. With a non-degenerate ground state, because $\Psi^{(2)}$ is not the ground state of $\hat{H}^{(1)}$,

$$E^{(1)} = \langle \Psi^{(1)} | \hat{H}^{(1)} | \Psi^{(1)} \rangle < \langle \Psi^{(2)} | \hat{H}^{(1)} | \Psi^{(2)} \rangle \tag{2.7}$$

and

$$\begin{aligned} \langle \Psi^{(2)} | \hat{H}^{(1)} | \Psi^{(2)} \rangle &= \langle \Psi^{(2)} | \hat{H}^{(2)} | \Psi^{(2)} \rangle + \langle \Psi^{(2)} | \hat{H}^{(1)} - \hat{H}^{(2)} | \Psi^{(2)} \rangle \\ &= E^{(2)} + \int d^3r \, [V_{ext}^1(\mathbf{r}) - V_{ext}^2(\mathbf{r})] \, n_0(\mathbf{r}). \end{aligned} \tag{2.8}$$

Therefore,

$$E^{(1)} < E^{(2)} + \int d^3r \, [V_{ext}^{(1)}(\mathbf{r}) - V_{ext}^{(2)}(\mathbf{r})] \, n_0(\mathbf{r}). \tag{2.9}$$

If the same steps are followed with $E^{(1)}$ instead of $E^{(2)}$, then

$$E^{(2)} < E^{(1)} + \int d^3r \, [V_{ext}^{(2)}(\mathbf{r}) - V_{ext}^{(1)}(\mathbf{r})] \, n_0(\mathbf{r}). \tag{2.10}$$

When Eq. 2.9 and Eq. 2.10 are combined, an impossible inequality is obtained: $E^{(1)}+E^{(2)}<E^{(1)}+E^{(2)}$. This demonstrates that the external potentials cannot be differing by more than a constant if they lead to the same non-degenerate ground state charge density. This proof can also be extended to the case of a degenerate ground state [54].

Theorem II: For any number of electrons N, a universal density functional for the energy $E[n]$ in terms of the density $n(\mathbf{r})$ exists such that the global minimum of $E[n]+\int d^3r\, V_{ext}(\mathbf{r})n(\mathbf{r})$ is the exact ground state energy of the system for any external potential $V_{ext}(\mathbf{r})$. The density that minimizes the functional is the exact ground state density $n_0(\mathbf{r})$.

Proof: Let us consider the HK total energy as a functional of the density

$$E_{HK}[n]=\underbrace{T[n]+E_{int}[n]}_{F_{HK}[n]}+\int d^3r\, V_{ext}(\mathbf{r})n(\mathbf{r})+E_{II}. \tag{2.11}$$

Here $F_{HK}[n]$ includes the internal energies of the electron system, and since it is a functional only of the density, it must be the same for all electron systems.

If a ground state density $n^{(1)}(\mathbf{r})$ of a system corresponding to the external potential $V_{ext}^{(1)}(\mathbf{r})$ is considered, the HK energy is the expectation value of the Hamiltonian $\hat{H}^{(1)}$ in the unique ground state with wave function $\Psi^{(1)}$ such that

$$E^{(1)}=E_{HK}[n^{(1)}]=\langle\Psi^{(1)}|\hat{H}^{(1)}|\Psi^{(1)}\rangle \tag{2.12}$$

Then, if a different density $n^{(2)}(\mathbf{r})$ is considered, the new energy must be larger than $E^{(1)}$, i.e.,

$$E^{(1)}=\langle\Psi^{(1)}|\hat{H}^{(1)}|\Psi^{(1)}\rangle<\langle\Psi^{(2)}|\hat{H}^{(1)}|\Psi^{(2)}\rangle=E^{(2)}. \tag{2.13}$$

This means that $E_{HK}[n]$ (Eq. 2.11) evaluated at the ground state density $n_0(\mathbf{r})$ is lower than the value of the same functional evaluated at any other density $n(\mathbf{r})$. Therefore, if the form of $F_{HK}[n]$ is known, the exact ground state density and energy can be found by minimizing the total energy with respect to changes in the density $n(\mathbf{r})$. However, the HK theorems do not give any insight into the exact form of the $F_{HK}[n]$.

Kohn and Sham

In the Kohn-Sham (KS) ansatz, it is assumed that the ground state density $n_0(\mathbf{r})$ of the original interacting system is equal to that of an auxiliary non-interacting system that can be solved more easily. This leads to the complicated many-body problem being replaced by an independent particle one, with the difficult many-body terms incorporated into an exchange-correlation (XC) functional.

Following the KS method, the auxiliary KS independent-particle system is determined by the auxiliary Hamiltonian with an effective potential:

$$\hat{H}^{\sigma}_{aux} = \sum_{i}^{N^{\sigma}} \left[-\frac{1}{2}\nabla_i^2 + V^{\sigma}_{eff}(\mathbf{r}_i) \right], \tag{2.14}$$

where N^{σ} is the number of electrons with spin σ. For a system of N electrons, the ground state has one electron in each of the spin orbitals $\psi_i^{\sigma}(\mathbf{r})$. Therefore the density of the auxiliary system is

$$n(\mathbf{r}) = \sum_{\sigma} n(\mathbf{r},\sigma) = \sum_{\sigma}\sum_{i=1}^{N^{\sigma}} |\psi_i^{\sigma}(\mathbf{r})|^2. \tag{2.15}$$

In the KS method, the Hohenberg-Kohn ground state energy functional (Eq. 2.11) is rewritten as

$$E_{KS} = T_s[n] + \int d\mathbf{r}\, V_{ext}(\mathbf{r}) n(\mathbf{r}) + E_{Hartree}[n] + E_{II} + E_{xc}[n], \tag{2.16}$$

where $T_s[n]$ is the independent-particle kinetic energy. $V_{ext}(\mathbf{r})$ is the external potential created by the nuclei and any other external fields, and $E_{Hartree}[n]$ is the classical Coulomb energy of the electron density interacting with itself. E_{II} is the interaction between nuclei, and $E_{xc}[n]$ is the XC energy, which groups together all of the many-body effects of exchange and correlation.

The exchange-correlation energy E_{xc} can be written as

$$E_{xc}[n] = \langle \hat{T} \rangle - T_s[n] + \langle \hat{V}_{int} \rangle - E_{Hartree}[n]. \tag{2.17}$$

In this form it is easy to see that the XC energy is the difference between the kinetic and internal energies of the interacting many-body system and those of the fictitious, independent-particle system with the same density. If the exact universal form of

$E_{xc}[n]$ were known, it could be used to solve the KS equations for an independent-particle system to find the exact ground state energy and density of the many-body system. However, the exact form is not known, and approximations to $E_{xc}[n]$ need to be used, as discussed in Section 2.3.

Once the form of $E_{xc}[n]$ is decided, the minimum of the Kohn-Sham energy can be found with the variation equation

$$\frac{\delta E_{KS}}{\delta \psi_i^{\sigma *}(\mathbf{r})} = \frac{\delta T_s}{\delta \psi_i^{\sigma *}(\mathbf{r})} + \left[\frac{\delta E_{ext}}{\delta n(\mathbf{r},\sigma)} + \frac{\delta E_{Hartree}}{\delta n(\mathbf{r},\sigma)} + \frac{\delta E_{xc}}{\delta n(\mathbf{r},\sigma)} \right] \frac{\delta n(\mathbf{r},\sigma)}{\delta \psi_i^{\sigma *}(\mathbf{r})} = 0, \tag{2.18}$$

with the constraint that the wave functions are orthonormal. This constraint is applied with Lagrange multipliers:

$$\delta \left[\langle \Psi | \hat{H} | \Psi \rangle - E(\langle \Psi | \Psi \rangle - 1) \right] = 0. \tag{2.19}$$

The resulting Kohn Sham equations are then

$$(H_{aux}^{\sigma} - \varepsilon_i^{\sigma})\, \psi_i^{\sigma}(\mathbf{r}) = 0, \tag{2.20}$$

where ε_i^{σ} are the eigenvalues and H_{aux}^{σ} is the Hamiltonian of Eq. 2.14, with

$$V_{eff}^{\sigma}(\mathbf{r}) = V_{ext}(\mathbf{r}) + \underbrace{\frac{\delta E_{Hartree}}{\delta n(\mathbf{r},\sigma)}}_{V_{Hartree}(\mathbf{r})} + \underbrace{\frac{\delta E_{xc}}{\delta n(\mathbf{r},\sigma)}}_{V_{xc}^{\sigma}(\mathbf{r})}. \tag{2.21}$$

The set of KS equations (Eq. 2.20) are independent-particle equations with an effective potential $V_{eff}^{\sigma}(\mathbf{r})$ and density $n(\mathbf{r},\sigma)$ that must be solved for self-consistently. Based on the HK theorems, the resulting ground state density then uniquely determines the KS effective potential $V_{eff}^{\sigma}(\mathbf{r})$ for each interacting electron system.

2.3 Exchange and Correlation Functionals

Although the exact form of the XC functional E_{xc} is not known, it can be approximated. Following the Kohn-Sham method, by explicitly separating the independent-particle energy and long-range Hartree terms of the total energy, the remaining part, which is the XC functional, can be approximated as a local or nearly local functional

of the density:

$$E_{xc}[n] = \int d\mathbf{r}\; n(\mathbf{r})\varepsilon_{xc}([n],\mathbf{r}). \tag{2.22}$$

Here $\varepsilon_{xc}([n],\mathbf{r})$, which describes the energy per electron at a point $\mathbf{r}$, is known as exchange and correlation energy density.

2.3.1 The Exchange and Correlation hole

In order to understand the physically motivated definition of $\varepsilon_{xc}([n],\mathbf{r})$, let us consider the concept of the XC hole. The correlation energy E_c can be defined as the difference between the exact energy and the energy of an uncorrelated state. If we consider the case of a pair of correlated electrons, the joint probability of finding an electron of spin σ at point $\mathbf{r}$ and one of spin σ' at point $\mathbf{r}'$ is $n(\mathbf{r},\sigma;\mathbf{r}',\sigma')$. However, for uncorrelated particles, the joint probability is just the product of probabilities $n(\mathbf{r},\sigma)$ and $n(\mathbf{r}',\sigma')$, and therefore the measure of correlation is

$$\Delta n(\mathbf{r},\sigma;\mathbf{r}',\sigma') = n(\mathbf{r},\sigma;\mathbf{r}',\sigma') - n(\mathbf{r},\sigma)n(\mathbf{r}',\sigma'). \tag{2.23}$$

This can be separated into the exchange and correlation parts, so that

$$\Delta n(\mathbf{r},\sigma;\mathbf{r}',\sigma') = n_{xc}(\mathbf{r},\sigma;\mathbf{r}',\sigma') = n_x(\mathbf{r},\sigma;\mathbf{r}',\sigma') + n_c(\mathbf{r},\sigma;\mathbf{r}',\sigma'). \tag{2.24}$$

Here $n_{xc}(\mathbf{r},\sigma;\mathbf{r}',\sigma')$ is referred to as the XC hole, which is a result of the inter-atomic repulsions caused by exchange and correlation interactions. An electron present at a point reduces the probability of finding one at $\mathbf{r}'$, creating a depletion, or hole, in the electron density around itself. One electron can only create a hole in the density equivalent to one missing electron. Consequently, to obtain physical accuracy, the exchange-correlation hole must obey the sum rule, which states that the integral over all $\mathbf{r}'$ must be equal to -1 per electron, i.e,

$$\int d\mathbf{r}'\; n_{xc}(\mathbf{r},\sigma;\mathbf{r}',\sigma') = -1. \tag{2.25}$$

2.3.2 The Adiabatic Approximation

An important simplification in DFT is its use of the adiabatic connection [39], which connects the interacting system that has the energy and density we are trying to find, to a fictitious, non-interacting system.

In the adiabatic connection, the electron Hamiltonian is

$$\hat{H}_\lambda = \hat{T} + \lambda \hat{V}_{int} + \hat{V}_{ext}, \tag{2.26}$$

where λ determines the amount of interaction in the system, varying from 0 (non-interacting system) and 1 (fully interacting system). $\hat{V}_{ext}$ is an external potential and $\hat{V}_{int}$ is the electron-electron interaction. The derivative of the energy with respect to the parameter λ, according to the Hellman-Feynman theorem [40, 28], can be expressed as

$$\frac{\partial E}{\partial \lambda} = \langle \Psi_\lambda | \frac{\partial \hat{H}_\lambda}{\partial \lambda} | \Psi_\lambda \rangle, \tag{2.27}$$

where Ψ_λ is the wave function for intermediate values of the interaction determined by λ. The finite energy difference between the interacting and non-interacting states is then

$$\Delta E = \int_0^1 d\lambda \frac{\partial E}{\partial \lambda} = \int_0^1 d\lambda \langle \Psi_\lambda | \frac{\partial \hat{H}}{\partial \lambda} | \Psi_\lambda \rangle. \tag{2.28}$$

Using the adiabatic connection, the interaction Hamiltonian is scaled by λe^2, so that the electron charge is varied from 0 to 1 in atomic units, with the constraint that the density is kept constant. Then it can be shown **??** that the XC energy, which is the difference between the interacting and non-interacting systems, is

$$E_{xc}[n] = \int_0^1 d\lambda \langle \Psi_\lambda | \frac{dV_{int}}{d\lambda} | \Psi_\lambda \rangle - E_{Hartree}[n] = \frac{1}{2} \int d^3r n(\mathbf{r}) \int d^3r' \frac{\bar{n}_{xc}(\mathbf{r}, \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}, \tag{2.29}$$

where

$$\bar{n}_{xc}(\mathbf{r}, \mathbf{r}') = \int_0^1 d\lambda n_{xc}^\lambda(\mathbf{r}, \mathbf{r}'). \tag{2.30}$$

This is the average over the coupling-constant λ of the exchange-correlation hole $n_{xc}(\mathbf{r}, \sigma; \mathbf{r}', \sigma')$ that has been summed over parallel ($\sigma = \sigma'$) and antiparallel ($\sigma \neq \sigma'$) spin.

The exchange-correlation energy density in Eq. 2.22 can then be written as

$$\varepsilon_{xc}([n], \mathbf{r}) = \frac{1}{2} \int d^3r' \frac{\bar{n}_{xc}(\mathbf{r}, \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{2.31}$$

2.3.3 First rung: The Local Density Approximation

The Local Density Approximation (LDA) is the first rung of what is referred to as "Jacob's Ladder" of density functional approximations, which builds up from the Hartree approximation to the best possible chemical accuracy [29]. Solids, because their electron density tends to vary smoothly, can often be considered close to the homogeneous electron gas limit. The LDA takes advantage of this and considers the XC energy density to be the same at each point in space as in the case of a homogeneous electron gas with the same density, for which it has been calculated with high accuracy using Monte Carlo Methods [17]. Then

$$E_{xc}^{LDA}[n] = \int d^3 n(\mathbf{r}) \varepsilon_{xc}^{hom}[n(\mathbf{r})], \tag{2.32}$$

where $\varepsilon_{xc}^{hom}[n(\mathbf{r})]$ is the XC energy density for a homogeneous electron gas. However, this method does not work very well in molecular systems with a more rapidly changing density, and has not been employed in the work presented in this thesis.

2.3.4 Second rung: Generalized-Gradient Approximation

The second rung of Jacob's Ladder is the Generalized Gradient Approximation (GGA), which adds an additional dependence on the magnitude of the gradient of the density $|\nabla n(\mathbf{r})|$ at each point. As a result, it works better for systems that are further from a homogeneous electron gas.

The GGA form of the energy functional can be written as [97]

$$\begin{aligned} E_{xc}^{GGA}[n] &= \int d^3 n\ (\mathbf{r}) \varepsilon_{xc}[n(\mathbf{r}), \nabla n(\mathbf{r})] \\ &= \int d^3\ n(\mathbf{r}) \varepsilon_x^{hom}[n(\mathbf{r})] F_{xc}[n(\mathbf{r}), \nabla n(\mathbf{r})], \end{aligned} \tag{2.33}$$

where F_{xc} is the dimensionless XC enhancement factor and $\varepsilon_x^{hom}[n(\mathbf{r})]$ is the XC energy density of a homogeneous electron gas. The approximate form of the enhancement factor F_{xc} can be determined by applying physical constraints, such as those mentioned in Appendix A, or by optimizing parameters to a reference database. There are several different forms of F_{xc}, and one of the most widely used ones is that of Perdew, Burke, and Enzerhof (PBE) [98]. The PBE form of F_{xc}, as well as a

few other forms, differing based on which physical constraints they were designed to obey, will be mentioned later in this thesis.

A large portion of the work presented in this thesis deals with the exchange-only enhancement factor $F_x(s)$, which is expressed as a function of the reduced density gradient

$$s = \frac{|\nabla n|}{2k_F n}, \tag{2.34}$$

where $k_F = (3\pi^2 n)^{1/3}$ is the so-called Fermi wave vector. The form of $F_x(s)$ then determines how much the exchange energy varies from the homogeneous gas case for different values of s. In following chapters, a new method of parametrizing $F_x(s)$ for water systems and tests of the accuracy of new GGA exchange functional forms are discussed.

2.3.5 Beyond the second rung

Although functionals beyond the second rung have not been been applied to the work presented in this thesis, they can significantly increase accuracy (and computational cost), depending on the system they are applied to. Here brief descriptions of the higher rungs given.

Third rung: The next step above GGA XC is the meta-GGA, which adds a dependence on second derivative of the density $\nabla^2 n(\mathbf{r})$ and/or kinetic energy densities $\tau_\sigma = 1/2\sum_i |\nabla \psi_i(n)|^2$.

Fourth rung: At the fourth rung are the hybrids functionals, which add exact exchange, the exchange calculated with the Hartree-Fock functional E_x^{HF}, and thereby reduce the self-interaction error (Section 2.5).

Fifth rung: The fifth rung is the highest rung of Jacob's ladder, in which the functional has an added dependence on unoccupied (or virtual) Kohn-Sham orbitals. Most of these functionals aim to improve the correlation effects and include orbital-dependent, non-local correlation, such as second order Møller-Plesset perturbation theory (MP2) [87] or Random Phase Approximations (RPAs) [15] in hybrid functionals.

Beyond the fifth rung: Beyond Jacob's Ladder, there are other exchange-correlations with added complexity that obey extra constraints [121], and the development of improved and more efficient density functionals is an ongoing process.

2.4 van der Waals interactions

The interactions among permanent dipoles and induced dipoles make up the vdW interactions: (i) the forces between pairs of permanent dipoles (Keesom forces), (ii) the forces between permanent dipoles and the corresponding induced dipoles (Debye forces), and (iii) the forces between pairs of instantaneously induced dipoles (London dispersion forces). These interactions decay relatively slowly with respect to interatomic separation r. At short distances r, they are relatively weak when compared to chemical bonds, but because of their relative slowness of decay, they become more dominant at larger distances.

It is well known that GGA XC functionals perform poorly when dealing with van der Waals (vdW) forces, because they neglect the non-local correlations involving interactions between induced dipoles [136, 55, 108]. There are several popular methods of incorporating vdW interactions into DFT functionals. For the calculations presented in this thesis, they were included by adding a non-local correlation term to the XC density functional proposed by Dion *et al*. in 2004 [26], termed van der Waals density functional (vdW-DF). In this method, the correlation part $E_c[n]$ of the exchange-correlation functional $E_{xc}[n]$ can be separated into both a local part $E_c^0[n]$ and a part that includes non-local correlations $E_c^{nl}[n]$. Then the total XC energy can be expressed as follows:

$$E_{xc}[n] = E_x[n] + E_c[n] = E_x[n] + E_c^0[n] + E_c^{nl}. \tag{2.35}$$

The local part of the correlation $E_c^0[n]$ is approximated with LDA, and in the original vdW-DF functional $E_x[n]$ is approximated with the revPBE [135] treatment of exchange. However, other forms of exchange can be used here, and we experimented with various GGA approximations for exchange in our calculations.[1] The general expression for the non-local part $E_c^{nl}[n]$ is

$$E_c^{nl}[n] = \frac{1}{2}\int\int d^3r d^3r' n(\mathbf{r})\phi(\mathbf{r},\mathbf{r}')n(\mathbf{r}'). \tag{2.36}$$

[1] The choice of exchange energy used in combination with the vdW-DF treatment of correlation is critical; with some treatments of the exchange energy in this model, the binding distances and energies can lose accuracy. With a good choice of exchange energy, however, the inclusion of dispersion forces has been shown to perform well for a wide range of systems [53].

Here the kernel $\phi(\mathbf{r},\mathbf{r}')$ is a function that only depends on $\mathbf{r}$ and $\mathbf{r}'$ through two functions d and d':

$$d(\mathbf{r},\mathbf{r}') = |\mathbf{r}-\mathbf{r}'|q_0[n(\mathbf{r}),\nabla n(\mathbf{r})],$$
$$d'(\mathbf{r},\mathbf{r}') = |\mathbf{r}-\mathbf{r}'|q_0[n(\mathbf{r}'),\nabla n(\mathbf{r}')]. \quad (2.37)$$

Here $q_0[n(\mathbf{r}),\nabla n(\mathbf{r})]$ is a universal functional that can be expressed in terms of the XC energy density $\varepsilon_{xc}^{hom}[n(\mathbf{r})]$ and exchange energy density $\varepsilon_{x}^{hom}[n(\mathbf{r})]$ of a homogeneous electron gas:

$$q_0[n(\mathbf{r}),\nabla n(\mathbf{r})] = \frac{\varepsilon_{xc}^{hom}[n(\mathbf{r})]}{\varepsilon_{x}^{hom}[n(\mathbf{r})]}k_F - \frac{Z_{ab}}{9}s^2 k_F, \quad (2.38)$$

where Z_{ab}=-0.8491, s is the reduced density gradient, and k_F is the Fermi wave vector.

As usual, there is a cost for the additional accuracy obtained with this treatment of vdW interactions, as the double integral in Eq. 2.36 over all space is computationally expensive. However, an algorithm developed by Román-Pérez and Soler [111] reduces the computational cost by transforming the double real space integral into reciprocal space. This efficient treatment of vdW-DF was applied to the work presented in this thesis.

2.5 Self-interaction errors in DFT

Self-interaction errors (SIE) result from the spurious interaction of an electron with itself. In the Hartree-Fock method, the Hartree energy (the third term in Eq. 2.4) does not vanish even for a one-electron system due to SIE. Inclusion of all of the terms in the Hartree-Fock corrects this problem, because all of the Hartree self-interaction terms are exactly cancelled by the corresponding exchange self-interaction terms (last term in Eq. 2.4). However, many DFT approximations for the XC energy contain SIE, and this is believed to be one of the main causes of failure among these approximations.

As an example, let us consider the energy of a one-electron system. The conditions for a functional to be free of SIE for this system are [103]:

$$E_x[n] + E_{Hartree}[n] = 0 \qquad \text{and}$$
$$E_c[n] = 0, \tag{2.39}$$

where $n(\mathbf{r})$ is the one-electron density. Because $E_{Hartree}[n]$ is a fully non-local functional of the density, the first requirement is not exactly satisfied with functionals on the first three rungs of Jacob's Ladder. With these functionals, the exchange energy density and, thus, the exchange energy $E_x[n]$ only depend on the local density. Thus, the exchange potential vanishes where the density vanishes, and the asymptotic falloff of the exchange potential and of the density are both exponential. The Hartree potential, however, shows the typical asymptotic behavior of a Coulomb potential, falling off as $1/r$. Because the exchange potential decays too rapidly in comparison to the Hartree potential, they cannot fully cancel with each other, especially in the asymptotic region. In addition, the second requirement, which implies that the correlation energy must be zero for a one-electron density, can only be satisfied with functionals on the third rung or higher [126].

Self-interaction is a problem that arises not only in one-electron systems, but also in many-electron systems, and is unfortunately difficult to quantify in a many-electron system. The most popular method to correct the SIE is by using the Perdew-Zunger (PZ) correction [103], although there are several methods developed to do this. For the work presented in this thesis, we have incorporated SIE corrections in some of our functionals by including a 1-body energy correction term to study the effects of SIE on water systems.

2.6 The SIESTA method

The SIESTA method [118] is an efficient implementation of DFT (designed to scale linearly with the number of atoms N, an $O(N)$ method), following the Kohn-Sham scheme (Section 2.2). The SIESTA code was the main instrument used for the calculations presented in this thesis. The development of a density functional as an addition to this code, and an analysis of its performance will be discussed in later chapters of this thesis.

2.6.1 Pseudopotentials

The behavior of electrons in an atom differs depending on their proximity to the nucleus. The innermost (core) electrons are very close to the nucleus, where strong nuclear potentials compensate for their higher kinetic energies, such that they become very localized and have minimal overlapping with the electrons from other atoms. The largest contribution that they make is to the screening of the nuclear potential, which leads to changes in the wave functions of the valence electrons. When a chemical species has many electrons, the core electrons do not participate in bonding. The outermost (valence) electrons, however, have a higher probability of ending up in interatomic regions, where bonds are formed.

It can be assumed that, for the most part, the contribution of the core electrons to the total binding energy does not change much when isolated atoms are brought together. Thus, approximations can be made that distinguish the valence electrons, and electrons that are in between both regions, from the core electrons. This is done by substituting the core electrons with an effective potential that produces the same wave functions for the outer electrons as the original core electrons do. These types of effective potentials are called pseudopotentials, and the basic idea for them comes from work done by Fermi in 1934 [27]. The use of pseudopotentials is very important, because it results in a great reduction in computational cost and, if done correctly, gives very accurate results.

There are many different methods for constructing pseudopotentials [69, 110, 125, 8, 47, 37, 124]. The SIESTA method generally uses the Troullier-Martins pseudopotentials, which meet the conditions required to be considered "norm-conserving pseudopotentials" [124]. Their construction involves defining the Hamiltonian that includes all of the electrons of an atom in order to obtain the states that correspond to the core and outer electrons. A potential is then found which precisely reproduces the orbitals of the outer electrons, and the Hamiltonian is solved for again with this potential. The resulting eigenvalues are then used to define the pseudopotential that is to be used.

In SIESTA, the pseudopotentials are read in their semilocal form, in which there is a different radial potential $V_l(\mathbf{r})$ for each angular momentum l, from a data file. However, a non-local pseudopotential is necessary for a more efficient representation of the combined effect of the nucleus and core electrons. The fully non-local

form is obtained by using the Kleinman and Bylander (KB) representation [51]:

$$\hat{V}^{PS} = V_{local}(\mathbf{r}) + \hat{V}^{KB}, \tag{2.40}$$

where

$$\hat{V}^{KB} = \sum_{lm} \frac{|\delta\hat{V}_l \phi_{lm}\rangle\langle\phi_{lm}\delta\hat{V}_l|}{\langle\phi_{lm}|\delta\hat{V}_l|\phi_{lm}\rangle}. \tag{2.41}$$

Here m is the magnetic quantum number, $|\phi_{lm}\rangle$ is an eigenstate of the atomic pseudo-Hamiltonian, and $\delta\hat{V}_l(\mathbf{r}) = V_l(\mathbf{r}) - V_{local}(\mathbf{r})$ is the difference between the local $V_{local}(\mathbf{r})$ and semilocal $V_l(\mathbf{r})$ parts of the pseudopotential.

2.6.2 Basis sets

A basis set is a set of functions that are used to expand molecular orbitals. There are multiple ways to do this, and SIESTA uses Linear Combinations of Atomic Orbitals (LCAOs). Within a cutoff radius r_c, each atomic basis orbital consists of the product of a spherical harmonic Y_l^m with a radial function ϕ_{Iln}. For an atom I located at $\mathbf{R}_I$,

$$\phi_{Ilmn}(\mathbf{r}) = \phi_{Iln}(r_I) Y_l^m(\hat{\mathbf{r}}_I), \tag{2.42}$$

where $\mathbf{r}_I = \mathbf{r} - \mathbf{R}_I$, $r = |\mathbf{r}|$, and $\hat{\mathbf{r}} = \mathbf{r}/r$. The angular momentum is denoted by l and m, and n differentiates the wave functions with the same angular momentum. When there are several orbitals with the same angular dependence, the basis set is termed *multiple* $-\zeta$. The radial functions are determined by cubic spline interpolations from values at points on a fine radial mesh.

The basis orbitals are brought to zero outside of r_c smoothly by calculating the numerical eigenfunctions $\phi_l(r)$ of the atomic pseudopotential $V_l(r)$ for a change in energy $\delta\varepsilon_l$, chosen such that $\phi_l(r_c) = 0$. Because of this method, it is common to refer to an "energy shift" rather than a cutoff radius. The Kohn-Sham orbitals are then represented as a linear combination of the atomic basis orbitals ϕ_{Ilmn}, and the corresponding electron densities can be determined.

By increasing the number of wave functions n used for the same atomic orbital with the split method [24], which is standard in quantum chemistry, the basis is improved by giving the orbital more degrees of freedom and thereby improving the chemical accuracy. In a multiple-ζ orbital, the atomic orbital is divided into parts: (i) an inner, compact orbital that is a linear combinations of Gaussians, determined ei-

ther variationally or by fitting numerical atomic eigenfunctions, and (ii) one or more diffuse functions, that are "released" or "split" from the combination. The diffuse functions are very shallow Gaussian basis functions that are designed to more accurately represent the tail portion of the atomic orbitals.

In order to improve the basis further and achieve better converged results, additional orbitals that have angular momentum higher than that of the highest occupied orbital, called polarization orbitals, can be added. The addition of polarization orbitals allows the possibility of a non-symmetric charge distribution, which occurs during bond formation and charge polarization. Instead of using orbitals with a higher angular momentum, extra polarization orbitals can be found by applying a small electric field ε, and then solving the Schrödinger equation using first-order perturbation theory. A basis set with one or more polarization orbitals will have "plus polarization" or "+p" added to its name, such as double-ζ plus polarization (dζ+p), etc. Note that, although using split valence and adding polarization orbitals improves the accuracy of calculations, it comes with an added computational cost.

Chapter 3

Simulating Water with *Ab intio* Methods

Liquid water is arguably the most important substance for life, as well as for a number of challenges considered to be of huge scientific and technological importance. At the same time, despite its molecular simplicity, it is a liquid of surprising complexity, with tens of thermodynamic anomalies. This complexity stems from the coexistence of covalent, electrostatic, and dispersion interactions, that have very different magnitudes but also must maintain a subtle and critical balance with each other. In order to make sense of the delicate interplay of forces involved in water systems and manage to properly represent them in simulations, it is necessary to first understand the details of how water molecules interact with each other. Section 3.1 gives a brief description of the important intermolecular interactions that determine the behavior of liquid water, and Section 3.2 discusses important factors to consider when choosing a functional for liquid water simulations and reviews results from previous studies involving *ab initio* Molecular Dynamics (AIMD) simulations of liquid water with density functionals of varying accuracy.

3.1 Liquid Water Interactions and Structure

A single water molecule is composed of two hydrogen (H) atoms covalently bonded to one oxygen (O) atom. The oxygen atom has two occupied molecular orbitals or "lone pairs", which are positioned opposite to the H atoms and perpendicular to the molecular plane. In condensed phases of water, the dominant intermolecular

interaction between a molecule and its neighbors is the hydrogen bond (H-bond), which is an interaction that is strongly dependent on bond direction. In a H-bond, the lone pair orbitals of the oxygen atom, being negatively charged, attract the positively charged hydrogen atoms of neighboring molecules. A molecule that has a positively charged hydrogen pointed towards a lone pair of another molecule acts as a H-bond donor, and a molecule that has a lone pair pointed towards the hydrogen of another molecule acts as a H-bond acceptor. When many water molecules are brought together in a condensed phase, a H-bond network is formed. In ice, this network is made up of tetrahedral structural units, resulting from the directions of the lone pair orbitals and of the O-H covalent bonds (Fig. 3.1).

The tetrahedral structuring formed by H-bonds is what causes water to contract upon melting, in contrast to most liquids, which expand upon melting. The structure of ice has a lot of open space, because the molecules are held strongly in four straight tetrahedrally-oriented H-bonds. Upon melting, some of these bonds break or bend, and the structure undergoes a partial collapse. On the other hand, with most other solids, the extra movement available in the liquid phase requires more space and therefore their melting is accompanied by expansion.

In its liquid form, at a given instant in time, water approximately keeps the tetrahedral formations, which are partially kept in place by the H-bonds. Many studies of the structure of liquid water using x-ray scattering, neutron-scattering and computer simulations show that the tetrahedral structure is distorted from that of ice due to the increased thermal motion of the molecules, with H-bonds constantly breaking and reforming on a time scale of approximately 0.5 picoseconds (ps) [131].

In addition to H-bond forces, there are non-directional, long-range van der Waals (vdW) forces, due to the dipole-dipole fluctuations of the water molecules, which are also a contributing factor to the structure of water. The strength of the H-bond in liquid water is roughly 0.2 eV and is normally stronger than vdW interactions ($\sim$0.02 eV) but weaker than the strong covalent chemical bonds ($\sim$2 eV).

Because of the strength of the H-bonds, they are difficult to break, giving water a very high boiling temperature, melting point, and viscosity when compared to other liquids that are not H-bonded. Depending on the definition of H-bond used in structural analysis, the average number of H-bonds per molecule in molecular dynamics simulations is usually between 3 or 4 at ambient conditions [59], corresponding to the approximate tetrahedron configuration.

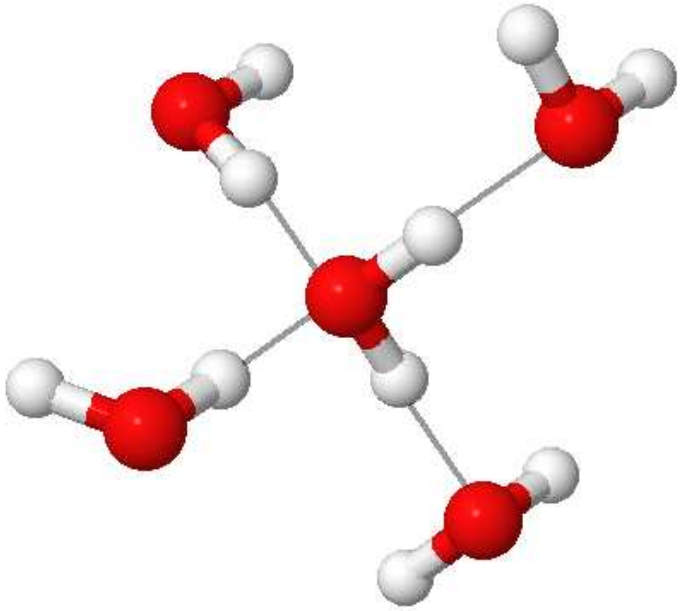


Fig. 3.1 Each water molecule, composed of one oxygen atom (red) and two hydrogen atoms (white) is part of the formation of a local tetrahedron in a H-bond network. The grey lines represent H-bond attractions between water molecules.

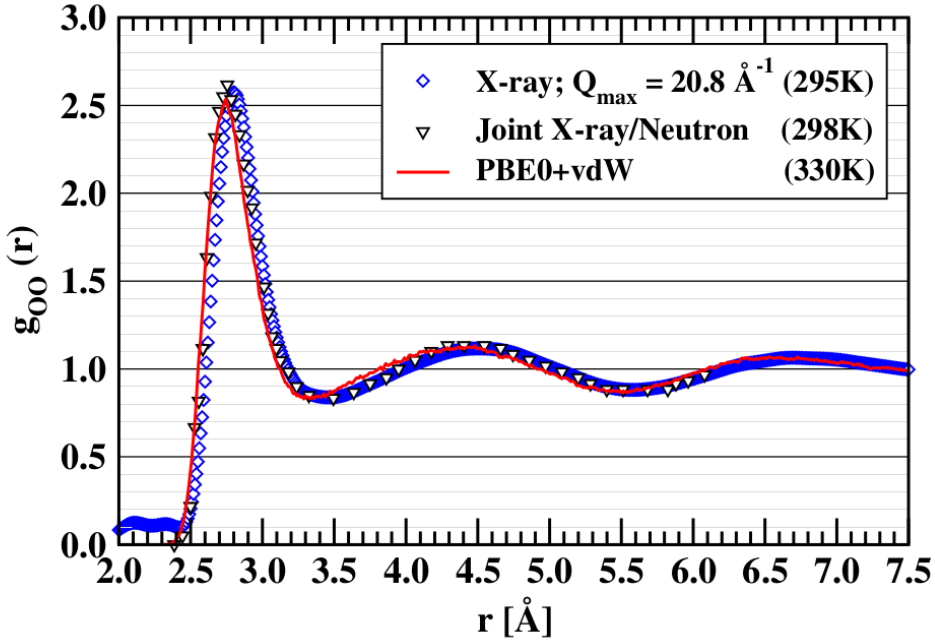
The distinction between what is and what isn't a H-bond between two water molecules is usually determined by the distance between the oxygen atoms of the molecules r_{OO} and the angle $\angle$O—H...O between the oxygen and hydrogen of the donor molecule and the oxygen of the acceptor, or by the distance r_{HL} between the hydrogen of the donor molecule and the center of the lone pair of the acceptor molecule. The cutoffs of these values, beyond which the molecules are not considered to be H-bonded, vary from study to study [84, 42, 106]. For all analyses presented in this thesis, two water molecules are considered to be H-bonded if the distance r_{HL} < 1.0 Å. The center of the lone pairs for each molecule are considered to be above and below the plane in which the oxygen and two hydrogen atoms lie, at a distance of 0.7 Å from the oxygen atom.

3.2 The Importance of vdW Interactions, Self-interaction Correction, and Nuclear Quantum Effects in Water

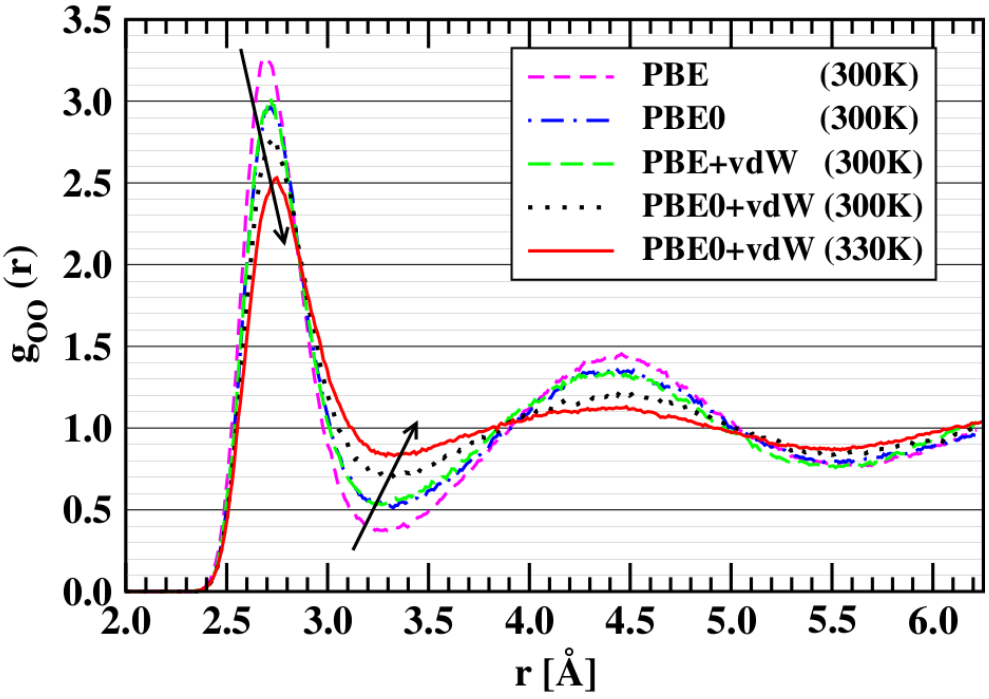
By gathering statistics about the properties of water, such as the structural and thermal properties of liquid water obtained from a simulation or the energetics of ice and water clusters, and comparing them with experiments, one can determine the degree of accuracy of DFT approximations for these systems. Many approximations have been tested in previous studies by performing *ab initio* Molecular Dynamics (AIMD) simulations and probing the liquid structure by analyzing the radial distribution functions (RDFs) extracted from the AIMD trajectories. In the following sections, the importance of vdW interactions, self-interaction corrections, and Nuclear Quantum Effects (NQE) in AIMD simulations of liquid water are discussed, and results from DiStasio, Jr. et al. [46] are shown (Fig. 3.2) as an example of the individual and collective influence of each of these effects.

3.2.1 vdW Interactions

Studies of liquid water usually focus on the oxygen-oxygen radial distribution function $g_{OO}(r)$, which gives important information about the liquid structure. The first peak in $g_{OO}(r)$ (also referred to as the first coordination peak or shell) gives information about the distribution of first neighbors that are H-bonded to each molecule. The second peak is produced by the next shell of molecules that are H-bonded. The height of the first minimum in $g_{OO}(r)$ reveals the population of the interstitial region between the first and second coordination peaks, which is formed by vdW interactions. Maintaining the proper balance between the populations of these regions is important. Peaks that are too high with respect to the population of the interstitial region would indicate that the liquid is over-structured, being too close to the rigid tetrahedral form seen in ice. Peaks that are too low with respect to the interstitial region reveal a liquid that is under-structured. To sustain the balance, it is clear that the correct relative magnitudes of H-bond and vdW interactions should be maintained. Thus, it is essential to include the proper description of vdW interactions in a functional used to study water.


X-ray; Q_{max} = 20.8 Å^{-1} (295K)
Joint X-ray/Neutron (298K)
PBE0+vdW (330K)
g_{OO} (r)
r [Å]

(a)


PBE (300K)
PBE0 (300K)
PBE+vdW (300K)
PBE0+vdW (300K)
PBE0+vdW (330K)
g_{OO} (r)
r [Å]

(b)

Fig. 3.2

Fig. 3.2 Taken from Ref. [46]. This figure shows the combined (a) and individual (b) effects of including vdW interactions, self interaction corrections, and nuclear quantum effects in liquid water simulations. In (a), the oxygen-oxygen RDF $g_{OO}(r)$ of liquid water obtained from an AIMD simulation at 300 K with PBE0+TS-vdW(SC), which takes into account all three effects, is shown to be in agreement with the experimental $g_{OO}(r)$ obtained from x-ray and scattering experiments [116, 18]. Here Q_{max} is the maximum value of Q, the change in the wavevector between the incident and scattered radiation, in the given scattering experiment. In (b), the $g_{OO}(r)$ of liquid water obtained from different DFT-based AIMD simulations are plotted. The arrows indicate the direction of the systematic shifts in the main RDF peak positions and intensities as the functionals are improved.

With the PBE functional, a widely used GGA functional that does not properly describe vdW interactions, H-bond interactions are overall too strong relative to the vdW interactions. As a result, an AIMD simulation of liquid water at ambient temperature with this functional results in a liquid that is over-structured, with the first and second peak of the oxygen-oxygen RDF $g_{OO}(r)$ too pronounced. When dispersion interactions are explicitly included by using the scheme proposed by Dion et al. [26], the structure of water softens, moving further from the rigid tetrahedron structure found in ice, with a lower average number of H-bonds per water molecule. The vdW forces, which are non-directional, strengthen non-H-bonded interactions, causing some H-bonds with the molecules in the first and second coordination shell to be broken and allowing more of them to populate the interstitial region. Thus, by encouraging more disordered configurations in the interstitial region, the inclusion of vdW forces lowers the first two peaks of $g_{OO}(r)$ and raises the minimum between them.

The effect of including vdW interactions can be seen in the results of DiStasio, Jr. et al. [46] in Fig. 3.2a. Here the resulting $g_{OO}(r)$ from a simulation done at 300 K with PBE is compared with that of PBE+TS-vdW(SC), which is the self-consistent (SC) dispersion-corrected analog of PBE, based on the Tkatchenko-Scheffler (TS) [123] density-dependent vdW functional [123]. Although the results of only this study are provided here as an example, similar improvements from GGA have also been found in other previous studies that included vdW density functionals [68, 129, 86].

Studies have also been done to test the effect of vdW interactions on ice systems. Results from studies done by Santra et al. [113] and by Wang et al. [124] showed that vdW interactions have a considerable effect on the transition pressures between crystalline ice phases. When DFT functionals that do not explicitly take vdW interactions into account are used, pressures are largely overestimated. This suggests that vdW interactions play an important role in defining the equilibrium density both in ice and liquid water, which has been confirmed in other studies [129, 115].

3.2.2 Self-interaction Errors

It has also been suggested that reducing self-interaction errors (SIE) is important when simulating water systems. By reducing the amount of self-interaction error in the exchange-correlation potential, the O-H covalent bonds become stronger and the intra-molecular O-H distances are shortened, bringing them closer to the experimental structure. This can lead to significant structural effects in water and other H-bonded systems, because it weakens the strength of H-bonds. It is well known that H-bonded systems show an anti-correlation effect [66] between the O-H covalent bond and the OH-O H-bond. As the O-H covalent bonds becomes stronger (they become shorter and their vibrational strength increases [81, 65]), the H-bonds weaken and result in larger OH-O distances.

Studies give mixed results about the effects of SIE corrections on the structuring of liquid water [46, 34, 12]. Some have shown no significant change, while others show a less-structured liquid water as a result of reducing SIE. By weakening the individual hydrogen bonds, a functional that reduces SIE allows a higher population of water molecules in the interstitial region. This effect can be seen in the results of DiStasio, Jr. et al. (Fig. 3.2b). They include the results of PBE0, which adds a correction to the SIE by mixing 25% of the exact exchange (Hartree–Fock exchange) with 75% of the PBE exchange [99, 49]. A comparison of the $g_{OO}(r)$ resulting from a PBE AIMD simulation at 300 K with that of PBE0 reveals a movement of the peaks in the same direction as seen with the inclusion of vdW forces. With PBE0, the first and second peaks become less pronounced and the first minimum is raised, indicating a lower population of water molecules in the coordination peaks in exchange for a higher population in the interstitial region.

Fig. 3.2b also shows the results from PBE0+TS-vdW(SC), which includes both the separate and combined effects of vdW interactions and SIE corrections. It can be seen that the inclusion of both effects brings $g_{OO}(r)$ even closer to the experimental results than the individual effect of including exact exchange or dispersion interactions does.

3.2.3 Nuclear Quantum Effects

Nuclear quantum effects (NQE) have a substantial impact on the behavior of water [92] and need to be considered in simulations to describe it properly. When the Born-Oppenheimer approximation is used, the movements of the electrons and nuclei are considered separately. In this approximation, the nuclei are treated classically and the electronic wavefunction for a specific nuclear composition is determined by solving the time-independent Shchrödinger equation for the electronic Hamiltonian. This approximation, however, does not properly include the effects of the quantum nature of the nuclei, such as zero-point vibrations and tunneling [57, 128, 80]. It is well known that these effects generally result in an increase in translational and rotational dynamics as well as a weakened H-bond network, and therefore give rise to a less-structured liquid [70, 91, 90, 79, 25, 36].

One way to explicitly include the quantum nature of the nuclei is to use path integral molecular dynamics (PIMD) simulations [84]. These calculations add a great deal of computational expense, although some studies involve schemes for estimating NQE that reduce the additional cost [120, 30]. In addition, when considering the RDFs involving oxygen atoms, an increase in temperature by 30 K in AIMD simulations has been shown to mimic the inclusion of NQE [84, 89].

The effects of the weakening of the H-bond network due to NQE are shown in the results of DiStasio, Jr. et al (Fig. 3.2). In addition to vdW effects and SIE corrections, the AIMD simulation at 330 K PBE0+TS-vdW(SC) takes NQE into account by increasing the temperature by 30 K with respect to the temperature of the other simulations. By mimicking the effect of the quantum nature of nuclei, the over-structuring of liquid water is further reduced, bringing it even closer to experiments. The combination of all three effects produces a $g_{OO}(r)$ that is closest to the experimental data (Fig. 3.2a).

3.2.4 Conclusions

In conclusion, vdW interactions, SIE corrections, as well as NQE have been shown to individually make a contribution to bringing the liquid water structure closer to that of experiments in comparison to standard GGA DFT methods. Therefore, to accurately study water systems through ab intio DFT simulations, it is important to include all of these effects. However, as usual in quantum chemistry methods, inclusion of each one of these effects typically adds to the computational cost in exchange for accuracy, depending on the implementation, and this can render the functional impractical for use in realistic simulations and calculations. Because water will continue to be an important system to study, it is worthwhile to continue working towards more efficient implementations of these effects.

In Chapter 6, results from AIMD simulations for liquid water will be presented. With all functionals used in the simulations, vdW interactions were explicitly taken into account in a non-local correlation part of the density functional, as explained in Chapter 2. Self-interaction corrections were included by correcting the 1-body energy term in some functionals. However, NQE effects have not yet been taken into account in the work presented in this thesis. In Chapter 6, our results are compared with those of a classical-nuclei simulation done with MB-pol, a potential which closely reproduces the experimental liquid structure of water when NQE are considered [76].

Chapter 4

The DPPS method and optimization of GGA exchange functionals for water

Parametrization and optimization methods are widely applied to design complex models, such as interatomic potentials. Most of these methods, however, are not straightforward, as they require choosing somewhat arbitrary functional forms that depend on many parameters and a complicated process of trial and error in their optimization. The number of parameters relative to the size of the reference dataset, while crucial to the outcome, depends on tricky and subjective decision making. We introduce an original and general optimization method, called Data Projection onto Parameter space (DPPS), that resolves much of this uncertainty. While the only application of DPPS presented in this thesis is the optimization of a GGA exchange functional to water energies, it is a general optimization method that can be applied to other functional forms, systems, and properties. In this chapter, the general method is introduced and illustrated through a few simple examples, and the specific application to the optimization of a GGA exchange functional for water is presented.

4.1 Methodology

4.1.1 Data Projection onto Parameter Space

The DPPS method is designed to find the best possible parameters used in a complex model that involves a scalar function, which depends on a large number of variables and parameters. As an example, consider the case of the optimization of a pairwise interatomic potential $V(R)$ of a single chemical species, in which the potential energy V is our scalar function and the atomic positions $\mathbf{R}_i$ are our variables on which it depends.

The first step of the application of DPPS to this example is to choose radial interpolation mesh points R_α for the interatomic distances. Then, the parameters of this model would be the values of the potential energy at the mesh points $v_\alpha \equiv V(R_\alpha)$. The interatomic potential can then be expressed as the interpolated function $V(R) = \sum_\alpha v_\alpha p_\alpha(R)$, where $p_\alpha(R)$ are an appropriate set of interpolation basis functions. These basis functions are determined unambiguosly by the interpolation scheme (e.g. cubic splines[1] [107]) and the interpolation points.

The next step is to choose a large reference dataset of known energies E^n_{ref} associated with system geometries $\mathbf{R}^n_i$, where i denotes atoms and n denotes systems or geometries. The reference energies can be obtained from experiments or from high accuracy calculations. We then would have everything we need to solve for the parameters that will best reproduce these energies, using the following equation:

$$E^n_{ref} = E^n \equiv \sum_{i<j} V(R^n_{ij}) = \sum_\alpha v_\alpha \sum_{i<j} p_\alpha(R^n_{ij}) \equiv \mathbf{v} \bullet \mathbf{g}^n, \tag{4.1}$$

where $\mathbf{v} \equiv \{v_\alpha\}$ and $\mathbf{g}^n \equiv \{g^n_\alpha\}$ are vectors in "parameter space", with $g^n_\alpha \equiv \sum_{i<j} p_\alpha(R^n_{ij})$ being proportional to the radial distribution function at the mesh distances R^n_{ij} between atoms i and j.

In Eq. 4.1, the values E^n_{ref} are the projections of the vector $\mathbf{v}$ of unknown parameters onto the vectors $\mathbf{g}^n$ of known data. Depending on the relative numbers of parameters and data, N_{dat} and N_{par}, respectively, the later will form a subspace of the former or they will overdetermine them. If $N_{dat} < N_{par}$, we can find the minimum vector $\mathbf{v}$ that satisfies Eq. 4.1 by setting $\mathbf{v} = \sum_n c_n \mathbf{g}^n$, substituting it into Eq. 4.1, and

[1] In the cubic splines method, it is also necessary to specify the two boundary conditions (either a given value of the first derivative, usually zero, or a zero second derivative).

solving for c_n. The straightforward result is $\mathbf{v} = \sum_{nm} E^n_{ref} S^{-1}_{nm} \mathbf{g}^m$, where $S_{nm} \equiv \sum_\alpha g^n_\alpha g^m_\alpha$. If $N_{dat} > N_{par}$, Eq. 4.1 can be solved with least-squares minimization, with the solution being $v_\alpha = \sum_\beta S^{-1}_{\alpha\beta} u_\beta$, where $S_{\alpha\beta} \equiv \sum_n g^n_\alpha g^n_\beta$ and $u_\beta \equiv \sum_n E^n_{ref} g^n_\beta$.

In practice, even if $N_{dat} >> N_{par}$, nearly all the data may lie in a subspace of the parameter space. As a result, certain combinations of parameters may be poorly determined, and very large parameter values will result if we require an exact fit of the data energies. This situation will be apparent by a nearly singular matrix $S_{\alpha\beta}$, with the standard remedy being to invert it by singular value decomposition, discarding the subspace with small eigenvalues. Another standard alternative is to add a penalty to large parameter changes, minimizing

$$Z = \sum_{n=1}^{N_{dat}} \left(\frac{E^n - E^n_{ref}}{\Delta E_n}\right)^2 + \gamma \sum_{\alpha=1}^{N_{par}} (v_\alpha - v^0_\alpha)^2 = \min, \tag{4.2}$$

where γ is a penalty parameter, v^0_α are an initial estimates of the parameter values and ΔE_n are the estimated errors in E^n and E^n_{ref}. In practice, γ can be defined as $1/(\Delta v)^2$, where Δv is the estimated error in the parameters. Here using γ is equivalent to the use of an eigenvalue cutoff in the singular value decomposition of $S_{\alpha\beta}$, but Eq. 4.2 provides a smoother and more natural transition to the Bayesian approach described in Section 4.1.3. It is important to note that, for a fixed value of γ, the resulting potential $V(R)$ is very insensitive to the number of interpolation parameters v_α, which can be arbitrarily large.

4.1.2 Functional optimization

Next consider the optimization of an XC energy as a functional of the electron number density $\rho(\mathbf{r})$ in the generalized gradient approximation (GGA):

$$E_{xc}[\rho(\mathbf{r})] = \int \rho(\mathbf{r})\ \varepsilon_{xc}(k_F(\mathbf{r}), k_G(\mathbf{r}))\ d^3\mathbf{r}, \tag{4.3}$$

where $k_F = (3\pi^2\rho)^{1/3}$ and $k_G = |\nabla\rho|/\rho$. To parametrize the XC energy density $\varepsilon_{xc}(k_F, k_G)$, we choose interpolation values of $k_{F\alpha}$ and $k_{G\beta}$ and define it at the interpolation points using $\varepsilon_{xc}(k_F, k_G) = \sum_{\alpha\beta}\ \varepsilon_{\alpha\beta}\ p_\alpha(k_F)\ p_\beta(k_G)$, so that $\varepsilon_{\alpha\beta}$ are our functional parameters. Assuming again that we have a dataset of system geometries $\mathbf{R}^n_i$ and corresponding total energies E^n_{ref}, we start with initial values $\varepsilon^0_{\alpha\beta} = \varepsilon^0_{xc}(k_{F\alpha}, k_{G\beta})$, where $\varepsilon^0_{xc}(k_F, k_G)$ is defined by some pre-existing reference GGA XC functional. We

find the self-consistent electron densities $\rho_n(\mathbf{r})$ and the initial total and XC energies, E_{tot}^{0n} and E_{xc}^{0n}, for each system n using the reference XC functional. Then, to the first order in $\delta\varepsilon_{\alpha\beta}$,

$$\begin{aligned}\delta E_{tot} &\simeq \sum_{\alpha\beta}\left(\frac{\partial E_{tot}}{\partial\varepsilon_{\alpha\beta}}+\int\frac{\delta E_{tot}}{\delta\rho(\mathbf{r})}\frac{\partial\rho(\mathbf{r})}{\partial\varepsilon_{\alpha\beta}}d^3\mathbf{r}\right)\delta\varepsilon_{\alpha\beta}\\ &= \sum_{\alpha\beta}\frac{\partial E_{xc}}{\partial\varepsilon_{\alpha\beta}}\delta\varepsilon_{\alpha\beta}\simeq\delta E_{xc},\end{aligned} \tag{4.4}$$

where we have used $\delta E_{tot}/\delta\rho(\mathbf{r})=0$ and $\partial E_{tot}/\partial\varepsilon_{\alpha\beta}=\partial E_{xc}/\partial\varepsilon_{\alpha\beta}$. Thus, we impose that the change in the total energy is equal to the change in the XC energy, i.e.,

$$E_{ref}^n-E_{tot}^{0n}=E_{xc}^n-E_{xc}^{0n}, \tag{4.5}$$

where E_{xc}^n is the XC energy that is needed to obtain E_{ref}^n, or the reference XC energy E_{ref-xc}^n. Then

$$\begin{aligned}E_{xc}^n=E_{ref}^n-E_{tot}^{0n}+E_{xc}^{0n}&=E_{ref-xc}^n\\ =\sum_{\alpha,\beta}\varepsilon_{\alpha\beta}\int\rho_n(\mathbf{r})p_\alpha(k_F(\mathbf{r}))p_\beta(k_G(\mathbf{r}))d^3\mathbf{r}&\\ =\varepsilon\bullet\rho^n&\end{aligned} \tag{4.6}$$

where $\varepsilon=\{\varepsilon_{\alpha\beta}\}$, $\rho^n=\{\rho_{\alpha\beta}^n\}$, and $\rho_{\alpha\beta}^n$ are the integrals in Eq. 4.6. The values of $\rho_{\alpha\beta}^n$ are the electron densities in parameter space, and they are closely related to the functions $g_1(r_s)$ and $g_3(s)$ of Zupan *et al* [137].

Although the values of $\varepsilon_{\alpha\beta}$ span a 2-D interpolation grid, rather than the 1-D grid of the value of v_α, Eqs. 4.1 and 4.6, they are entirely equivalent. Thus, we can solve Eq. 4.2, with $\varepsilon_{\alpha\beta}$ instead of v_α, to find a new set of parameters $\varepsilon_{\alpha\beta}$ (i.e. a new XC functional). We can then iterate to obtain selfconsistency between $\varepsilon_{\alpha\beta}$, E_{xc}^n, and $\rho(\mathbf{r})$.

For GGA exchange-only functionals, a simple scaling law requires the exchange energy density to be the product of a functional of a single variable, the reduced gradient $s\equiv k_G/2k_F$, and the exchange energy density of the homogeneous electron gas (HEG) (labelled LDA because of its use in the local density approximation):

$$\varepsilon_x(k_F,k_G)=\varepsilon_x^{LDA}(k_F)\,F_x(s), \tag{4.7}$$

with $\varepsilon_x^{LDA}(k_F) = -3k_F/4\pi$. Thus, as in the pair-potential example, in this case we can use a single variable interpolation for the unknown functional $F_x(s)$, usually called the "exchange enhancement factor."

In general, known experimental data refer to energy differences rather than total energies. Therefore we may use accurate reaction and atomization energies, or energy differences between different solid phases. To use these data efficiently, it suffices to substitute all energies E^n by ΔE^n and ρ_n by $\Delta\rho_n$ in Eq. 4.6, where Δ refers to the difference between the two systems. Then

$$\Delta E^n_{ref} - \Delta E^{0n}_{tot} + \Delta E^{0n}_{xc} = \sum_{\alpha,\beta} \varepsilon_{\alpha\beta} \int \Delta\rho_n(\mathbf{r}) p_\alpha(k_F(\mathbf{r})) p_\beta(k_G(\mathbf{r})). \tag{4.8}$$

This procedure also allows one to use a large variety of structural and thermodynamic data. For example, to impose known equilibrium geometries, we consider two systems with one of the atoms displaced by $\pm\Delta R$. Since the force must be zero at the equilibrium geometry, $\Delta E^n = 0$. The same can be done to impose zero pressure and stress at the known equilibrium geometry of a solid.

Equally, known vibration frequencies ω_q can be imposed using frozen-phonon displacements. The polarization vectors u_{iq}, where i denotes the atomic coordinates and q is a phonon index, can be calculated using an initial functional. The "experimental" Hessian is $H_{ij} = \sum_q \sqrt{m_i m_j}\omega_q^2 u_{iq} u_{jq}$, where ω_q are the experimental frequencies, which can then be imposed by displacing two coordinates and setting

$$\begin{aligned} H_{ij} = (&E^n_{tot}(R_i + \Delta R, R_j + \Delta R) - \\ &E^n_{tot}(R_i + \Delta R, R_j - \Delta R) - \\ &E^n_{tot}(R_i - \Delta R, R_j + \Delta R) - \\ &E^n_{tot}(R_i - \Delta R, R_j - \Delta R))/\Delta R^2, \end{aligned} \tag{4.9}$$

which defines a set of "experimental" changes in the XC energy that can be solved for using Eq. 4.8.

Alternatively, a larger set of random geometries and corresponding energies (possibly taken from a molecular dynamics simulation) can be used to optimize both the equilibrium geometry and the deformation energies, beyond the harmonic approximation. While DPPS is a general method that can be applied to many types of data, so far we have only tested its application to energies of random geometries, as presented in following sections.

4.1.3 Bayesian constraints

Although the simple penalty term of Eq. 4.2 avoids large parameter changes, in regions that are poorly determined by the data, it contains very little of our *ab initio* knowledge (and uncertainty) of how the functional should be. Therefore, the resulting functional, though optimal to fit a restricted set of data, will in general be rather unplausible from a theoretical point of view, and unreliable to reproduce other data. An obvious solution would be to minimize the error in the calculated energies, as in Eq. 4.2, but under a number of specific theoretical constraints, not just a general penalty term. Some of those constraints should be strict, while others may be more "relaxed" (quantitative). Thus, we know for sure that $F_x(0) = 1$, but we are much less certain about high values of s. More generally, the problem is to encode efficiently the known theoretical information, either strict or ambiguous, and this is exactly the aim of Bayesian probability theory [13].

Bayes theorem can be summed up in the following equation:

$$\mathscr{P}(theory|\,facts) = C\ \mathscr{P}(\,facts|theory)\mathscr{P}(theory), \tag{4.10}$$

where $\mathscr{P}(theory|\,facts)$ is the probability (or likelihood) that a theory is true, given that some facts have been observed, $\mathscr{P}(\,facts|theory)$ is the probability that those facts would be observed if the theory was true, $\mathscr{P}(theory)$ is (our estimate of) the *a priori* probability that the theory is true, and C is a normalization constant. In our case, "theory" means a quantitative parametrization of a given functional form (e.g. a set of GGA values $\varepsilon_{\alpha\beta}$), "facts" are a set of E^n_{ref} and E^n_{GGA} energies, and "true" means optimal to reproduce the energies (not only of our dataset, but of all possible systems of interest).

Assuming gaussian probability distributions,

$$\mathscr{P}(\,facts|theory) = C_1 \exp\left[-\sum_n \frac{1}{2}\left(\frac{E^n_{GGA} - E^n_{ref}}{\Delta E_n}\right)^2\right], \tag{4.11}$$

where ΔE_n are the estimated errors in the computed energies, due to causes not related to the XC functional (e.g. basis set incompleteness in E^n_{GGA}, or experimental errors in E^n_{ref}), and to the inability of the GGA functional form to reproduce the exact energies.

Equally,

$$\mathscr{P}(theory) = C_2 \exp[-\sum_{\alpha\beta} \frac{1}{2}(\varepsilon_\alpha - \varepsilon_\alpha^0) Cov_{\alpha\beta}^{-1} (\varepsilon_\beta - \varepsilon_\beta^0)], \tag{4.12}$$

where ε_α^0 is the average of $\varepsilon_{xc}(k_{F\alpha}, k_{G\alpha})$ among different *ab initio* GGA functionals:

$$\varepsilon_\alpha^0 = \sum_i w_i \varepsilon_\alpha^i, \tag{4.13}$$

where index i labels different functionals [102, 10, 61, 98, 135, 38, 132, 100, 74, 19, 14], w_i are normalized weights assigned to each one, and $\varepsilon_\alpha^i \equiv \varepsilon_{xc}^i(k_{F\alpha}, k_{G\alpha})$. In order to simplify the notation, and in contrast with previous paragraphs, here we are using a single index α for each pair of values $(k_F, k_G)_\alpha$, despite the fact that these pairs form a 2-D mesh (or a higher-dimensional mesh in case of a more complicated functional form, like meta-GGAs). $Cov_{\alpha\beta}^{-1}$ is the inverse of the covariance matrix

$$Cov_{\alpha\beta} = \sum_i w_i (\varepsilon_\alpha^i - \varepsilon_\alpha^0)(\varepsilon_\beta^i - \varepsilon_\beta^0). \tag{4.14}$$

Note that, in Eq. 4.12, we are using the discrepancies between the different *ab initio* functionals as a measure of our uncertainty of its exact form. However, the term $\mathscr{P}(theory)$ in Eq. 4.10 is important mostly in regions of the functional domain that are poorly sampled by the data, or in which the theoretical constraints are strict.

In practice, the estimated errors ΔE_n in Eq. 4.11 can be used partially as a knob to balance our relative uncertainties of the reference and DFT calculations, and also to take into account the uncertainty of the *ab initio* functional. Also, note that setting $Cov_{\alpha\beta} = \frac{1}{\gamma} I_{\alpha\beta}$ makes Eq. 4.10 equivalent to Eq. 4.2. In the following sections, Eq. 4.2 is referred to as the *unconstrained* form of DPPS, while Eqns 4.10-4.14 are referred to as the *constrained* form.

4.2 Optimization of a GGA exchange functional for water

As an example of an application of the DPPS methodology, we optimize a GGA exchange functional to a dataset of water energies predicted by high accuracy quantum chemistry methods. Given the narrow scope of our system, we cannot assure that the resulting functionals will be significant in a broader range of DFT applica-

tions. Rather, our aim is to shed new light on why DFT has been so frustratingly poor in simulating liquid water, on what changes are needed to improve it, and on the intrinsic shortcomings of the functional form for this important system.

Previous studies have found that for water systems, n-body interactions with $n \geq 4$ are relatively small and therefore it is not necessary to correct these terms in order to achieve chemical accuracy [35, 133, 96, 43, 134, 22, 41, 48]. It has also been shown that, by using functionals that include nonlocal vdW interactions and a hybrid functional for exchange to correct self-interaction errors, the overstructuring in radial distribution functions of liquid water produced by GGA based on ab-initio molecular dynamics (AIMD) simulations can be reduced [112, 46]. By optimizing the GGA exchange of the vdW-DF-cx [14] functional with DPPS to better reproduce water system energies up to 3-body terms, our aim is to obtain similar improvements with a functional that is less computationally expensive than those including hybrid exchange functionals.

In this section, we use DPPS to analyze the changes needed in the GGA exchange functional for the best description of water. First, the modification of the $F_x(S)$ with unconstrained and constrained DPPS to fit to 1-, 2-, and 3-body reference energies, separately, is explored. Functionals optimized to both 2- and 3-body terms simultaneously are then studied. The unconstrained versions of these functionals allow us to see the best possible functional form for reproducing the reference dataset, but they disregard physical plausibility. The constrained versions are the more practical part of the solution that can be applied to calculations involving larger systems of water.

4.2.1 Computational details

We optimize the GGA exchange starting from the vdW-DF-cx functional [14], which is an exchange-modified version of the original vdW-DF [68] functional. The vdW-DF functional uses the revPBE [135] form of $F_x(s)$ in the GGA desciption of exchange. The vdW-DF-cx functional, however, uses a $F_x(s)$ that starts with the form of Langreth-Vosko (LV) [60] in the region where $s < 2$ and transitions to the revised Perdew-Wang-86 (PW86R) form [102, 85] form for larger values of s. In the following, the exchange enhancement factor of the vdW-DF-cx functional is therefore labelled LV-PW86R and is plotted in Fig. 4.1 along with the $F_x(s)$ of the LV and

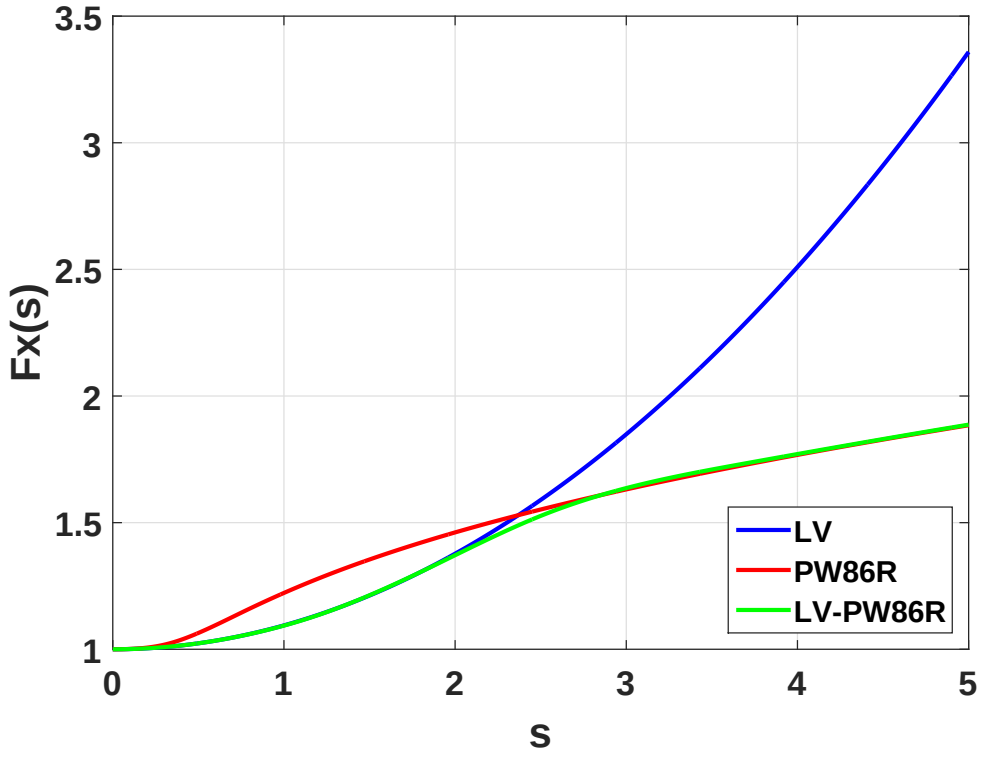


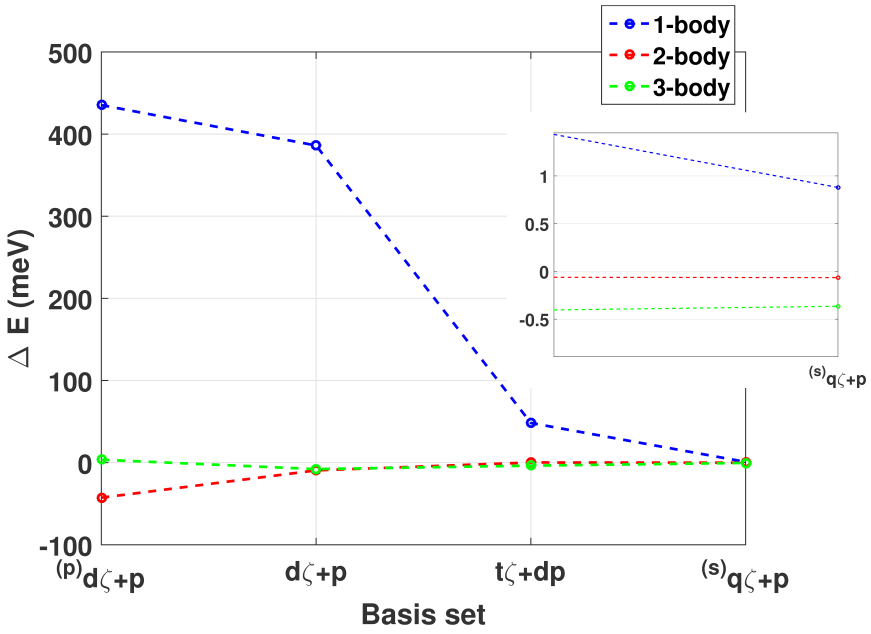
Fig. 4.1 The GGA exchange enhancement factors $F_x(s)$ of the LV (blue), PW86R (red), and LV-PW86R (green) exchange functionals.

PW86R functionals. The LV-PW86R exchange enhancement factor is

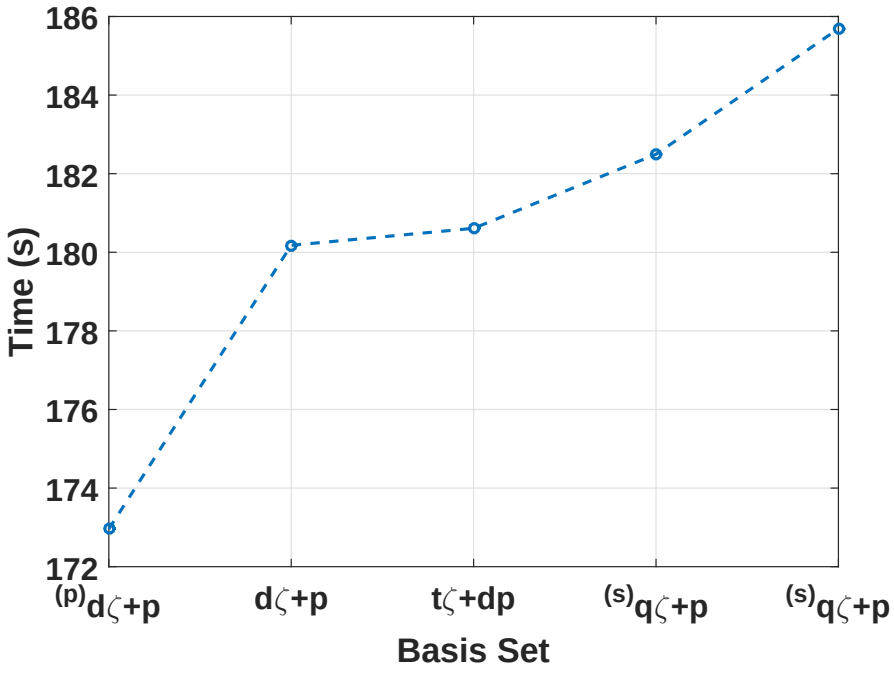
$$F_x^{LV-PW86R}(s) = \left(\frac{1}{1+\alpha s^6}\right) F_x^{LV}(s) + \left(\frac{\alpha s^6}{\beta+\alpha s^6}\right) F_x^{PW86R}(s), \tag{4.15}$$

where $\alpha = 0.02178$ and $\beta = 1.15$. Here $F_x^{LV}(s) = 1+\mu_{LV} s^2$ is the LV form with $\mu_{LV} = Z_{ab}/9$ and $Z_{ab} = -0.8491$, and $F_x^{PW86R}(s) = (1+15as^2+bs^4+cs^6)^{1/15}$ is the PW86R form, with $a = 0.1234$, $b = 17.33$, and $c = 0.1630$.

For an initial application of the DPPS method, we optimized the first three terms of the many-body expansion of the total energy of water, which was introduced in Eq. 1.2 of Chapter 1. The first term is the sum of the 1-body energies, which are the energies of the individual water monomers. The second term is the 2-body energy, which is the total energy of a dimer, relative to the two separated monomers in their same geometries. According to Eq. 1.2, the 2-body energy of each dimer is

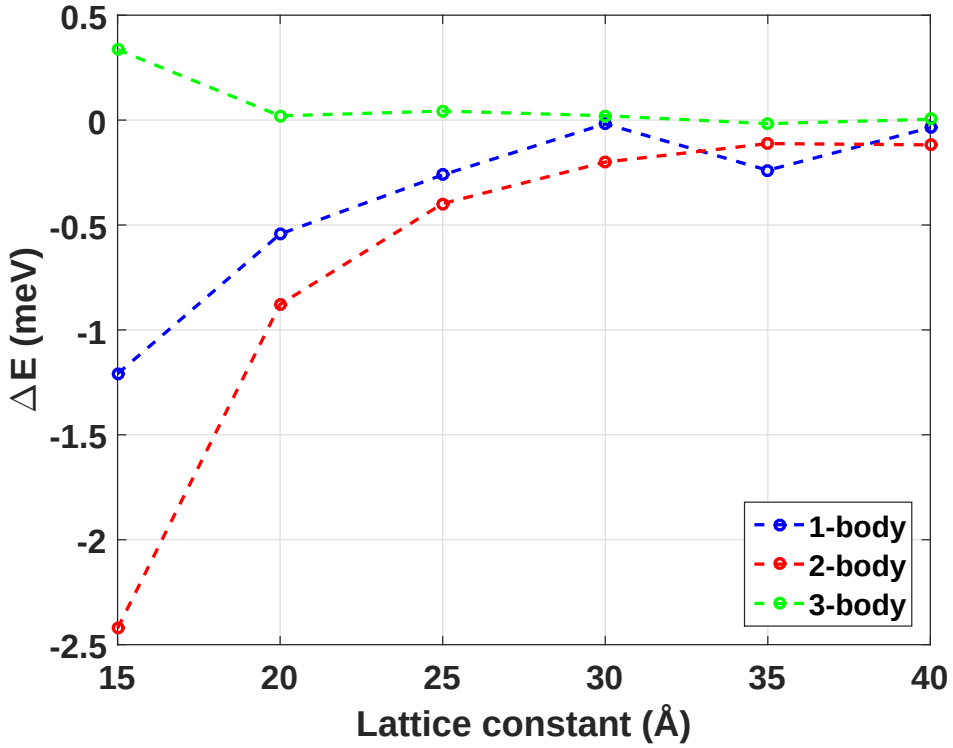


(a)

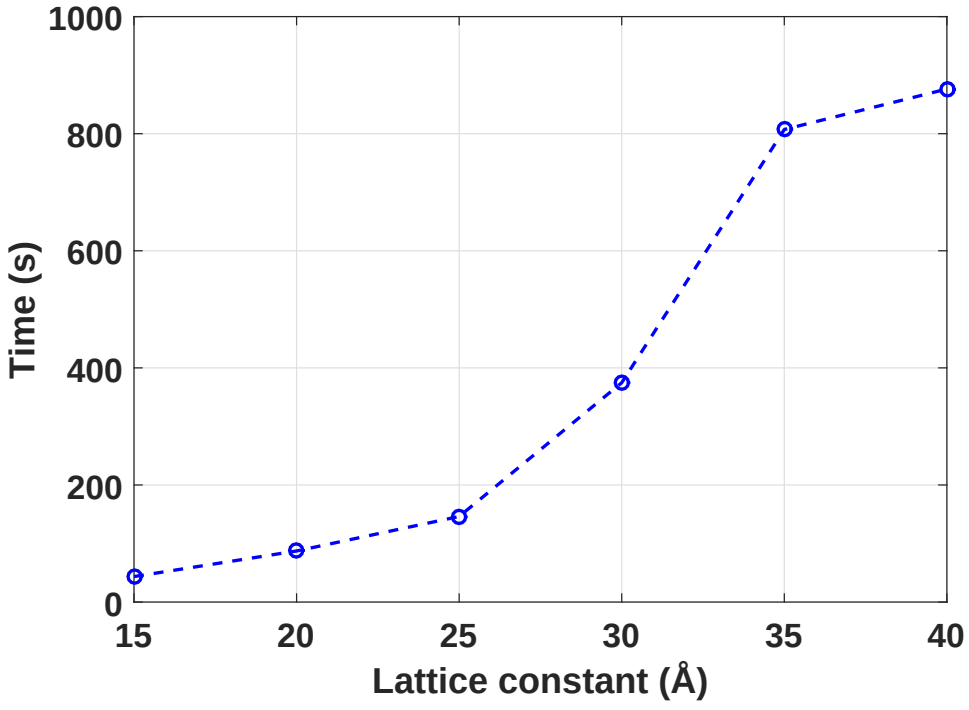


(b)

Fig. 4.2 Convergence test of 1-,2-, and 3-body energies with the ${}^{(p)}d\zeta$+p, dζ+p, tζ+dp, and ${}^{(s)}q\zeta$+dp basis sets. Plotted here are (a) the difference of the total energies (ΔE) with respect to qζ+dp and (b) the wall clock times of the monomer energy calculations for one SCF iteration with different bases. A 20 Å lattice constant and a 300 Ry mesh cutoff were used. The inset shows the values for ${}^{(s)}q\zeta$+dp.

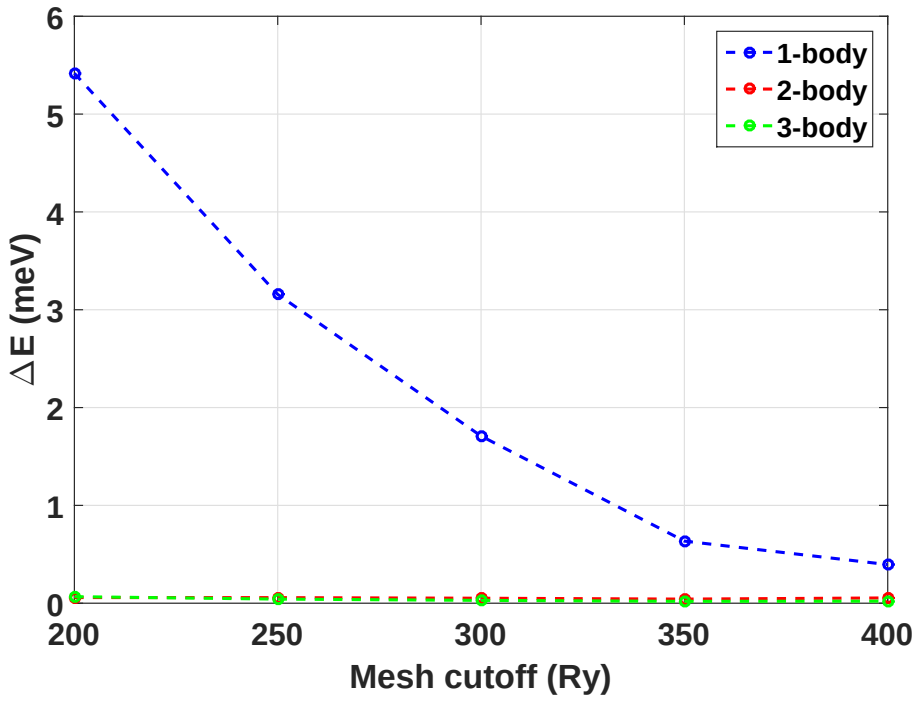


(a)

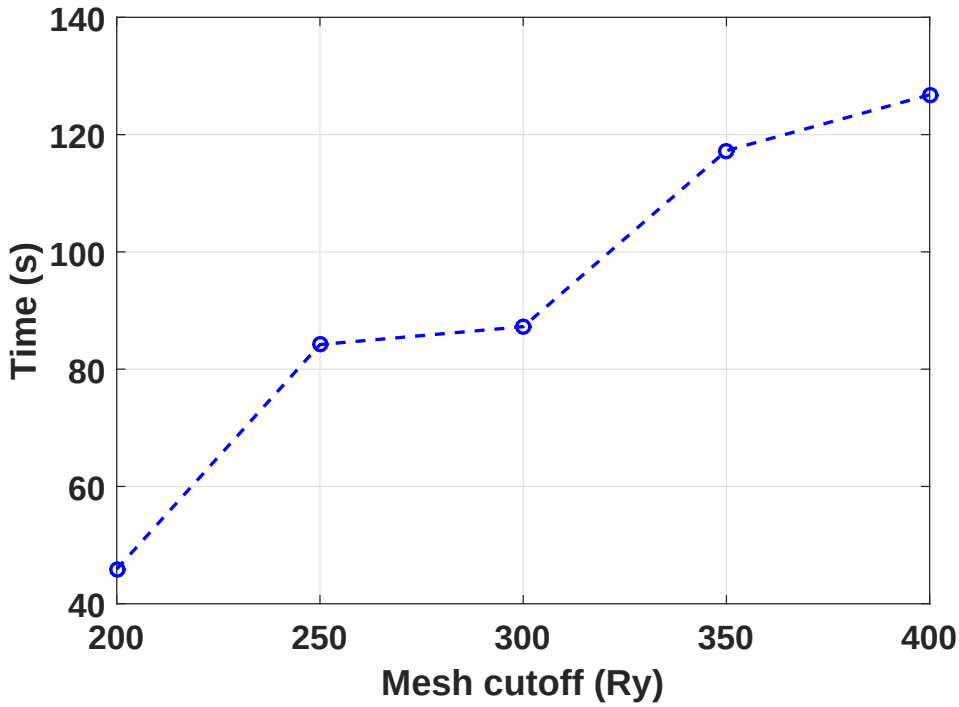


(b)

Fig. 4.3 Plotted here are (a) the errors of 1-,2-, and 3-body energies with respect to those calculated with a 50 Å lattice constant and (b) the calculation times for one SCF iteration of a monomer energy calculation with different lattice constants. A 300 Ry mesh cutoff and the $^{(s)}q\zeta$*+dp basis set were used.*



(a)



(b)

Fig. 4.4 Plotted here are (a) the errors of 1-,2-, and 3-body energies with respect to a those calculated with a 500 Ry mesh cutoff and (b) times for one SCF iteration with different mesh cutoffs in the calculation of the energy of a water monomer. A 20 Å lattice constant and $^{(s)}q\zeta$*+dp basis set were used.*

expressed as

$$V^{(2B)}(x_1,x_2) = V_2(x_1,x_2) - \sum_{a=1}^{2} V^{(1B)}(x_a), \tag{4.16}$$

where x_a are the coordinates of all the atoms in the a^{th} molecule, $V_2(x_1,x_2)$ is the total dimer energy, and $V^{(1B)}(x_a)$ are the energies of the monomers in the dimer. The third term of the many-body expansion is a sum of the 3-body energies, which can be expressed as

$$V^{(3B)}(x_1,x_2,x_3) = V_3(x_1,x_2,x_3) - \sum_{a=1}^{3} V^{(1B)}(x_a) - \sum_{a>b} V^{(2B)}(x_a,x_b), \tag{4.17}$$

where $V_3(x_1,x_2,x_3)$ is the total energy of a trimer, $V^{(1B)}(x_a)$ are the energies of the monomers in the trimer, and $V^{(2B)}(x_a,x_b)$ are the 2-body energies for all of the possible dimer combinations in the trimer.

Before calculating the initial energies and densities for each energy term, convergence tests were performed to find a good balance between accuracy and computational cost for the following simulation parameters:

- *Basis set*. Different types of basis sets, and how they affect the convergence of results is discussed in Section 2.6.2.
- *Lattice constant*. This defines the value of a, which determines simulation boxes of dimensions a x a x a.
- *Mesh cutoff*. This determines the fineness of the real-space grid used for integration.

The convergence of the 1-, 2-, and 3-body energies of a water monomer, dimer, and trimer, respectively, with the implementation of the vdW-DF-cx functional in a DPPS-enhanced version of the SIESTA code [118] was tested with the following basis sets [21]:

- $^{(p)}d\zeta$+p.[2] A double-ζ basis set with a single polarization shell and a cutoff radius $r_c = 4.5$ Bohr, beyond which the radial part of the basis function are strictly zero.
- $d\zeta$+p. A double-ζ basis set with a double polarization shell and $r_c = 8.5$ Bohr (dζ+p).

[2] p denotes an earlier basis set first proposed in Ref. [129]

- *t*ζ+*dp*. A triple-ζ basis set with a double polarization shell and $r_c = 8.5$ Bohr.
- $^{(s)}$*q*ζ+*dp*.[3] A quadruple-ζ basis set, with a double polarization shell and $r_c = 8.5$ Bohr.

The errors $\Delta E = E_{calc} - E_{ref}$ of the calculated energies E_{calc} are relative to the reference energies E_{ref}. For the basis set convergence test, E_{ref} were obtained with a quadruple-ζ basis set, with a double polarization shell and $r_c = 10$ Bohr (qζ+dp). Energy errors with these different basis sets are plotted in Fig. 4.2a and the time to complete one self-consistent field (SCF) iteration of a monomer energy calculation for each basis set are plotted in Fig. 4.2b. Because all of the energies are well converged with a $^{(s)}$qζ+dp basis, with a computation time insignificantly higher than with the tζ+dp basis set, we chose this basis set for all DFT total energy calculations.

The convergence of the energies with different lattice constants was also tested (Fig 4.3). For this test, the values of E_{ref} were calculated with a 50 Å lattice constant. The calculation time was found to be quite sensitive to the lattice constant (Fig. 4.3b), and a value of 20 Å was found to give a reasonable balance between accuracy and computational cost.

For the test of the mesh cutoff values (Fig. 5.7), the values of E_{ref} were calculated a 400 Ry mesh cutoff. A mesh cutoff of 300 Ry was determined to show the best balance between accuracy and computational cost for our calculations. Therefore, unless otherwise noted, all of the calculations presented in the remainder of the chapter were performed with a $^{(s)}$qζ+dp basis set, a 20 Å lattice constant, and a 300 Ry mesh cutoff.

4.2.2 Test: Does DPPS find the optimal GGA exchange functional?

Before performing a more comprehensive fit of a GGA exchange functional for water, as a test of the DPPS method, we used it to reproduce the $F_x(s)$ of PW86R [102, 85], starting from the $F_x(s)$ of PBE [98] by optimizing to PW86R energies. If the optimized form of $F_x(s)$ is the same as that of PW86R, this means that DPPS is able to find the exact GGA exchange functional form needed to obtain a set of reference (PW86R) exchange energies.

[3] s denotes a basis set with a short cutoff radius r_c

For this test, we used a set of 198 water dimer configurations, which were obtained by drawing dimers out of a simulation of the liquid performed with the classical AMOEBA force field of Ren and Ponder [105]. The dimers were drawn at random, but with the restriction that the O-O distance had to be less than 4.5 Å. The reference energies E^n_{ref} used here are the total dimer energies calculated with PBE correlation and PW86R exchange. The input energies E^n and densities $\rho(r)$ were calculated using PBE exchange and correlation.

Fig. 4.5 shows the initial values of $F_x(s)$ (PBE) plotted along with the reference $F_x(s)$ (PW86R) and values of the exchange enhancement factor $F_{x\alpha}$ at the interpolation points s_α=0.00, 0.0439, 0.107, 0.198, 0.328, 0.516, 0.787, 1.18, 1.73, 2.54, and 3.70, optimized to reproduce the PW86R exchange energies of the dimers (labelled CUSTOM). These CUSTOM values of $F_x(s)$ were obtained by minimizing the target function Z of Eq. 4.2, with $F_{x\alpha}$ in place of v_α:

$$Z = \sum_{n=1}^{N_{dat}} \left(\frac{E^n - E^n_{ref}}{\Delta E_n}\right)^2 + \sum_{\alpha=1}^{N_{par}} \left(\frac{F_{x\alpha} - F^0_{x\alpha}}{\Delta F}\right)^2 = \min. \tag{4.18}$$

Here $F^0_{x\alpha}$ and $F_{x\alpha}$ are the initial and final enhancement factor values, respectively and ΔF is the estimated error in the enhancement factor. For this test, we used $\Delta F = 2$ and $\Delta E_n = 0.1$ meV for all dimer energies. 20 interpolation points s_α between 0 and 100 were used to define the parameters $F_{x\alpha}$, with a greater density of interpolation points at lower values of s. The same interpolation points were used in all of the exchange functional optimizations presented in this chapter. Eq. 4.18 is minimized for all of the following *unconstrained* fits of $F_{x\alpha}$ referred to in this chapter.

A comparison of the forms of $F_x(s)$ plotted in Fig 4.5 reveals that DPPS moves the values of $F_{x\alpha}$ correctly so that the CUSTOM values line up with the PW86R form of $F_x(s)$. This demonstrates that DPPS is able to do exactly what we want it to: it finds the specific required GGA exchange functional necessary to reproduce a set of reference energies. Note that the input densities $\rho(r)$, obtained with PBE exchange, were left unchanged in this test, indicating that the errors in $F_x(s)$ introduced by using the initial $\rho(r)$ are negligible.

4.2.3 Dataset details

As a step towards obtaining a better description of liquid water, we have applied DPPS to the analysis and optimization of the energies of 300 water monomers, 330

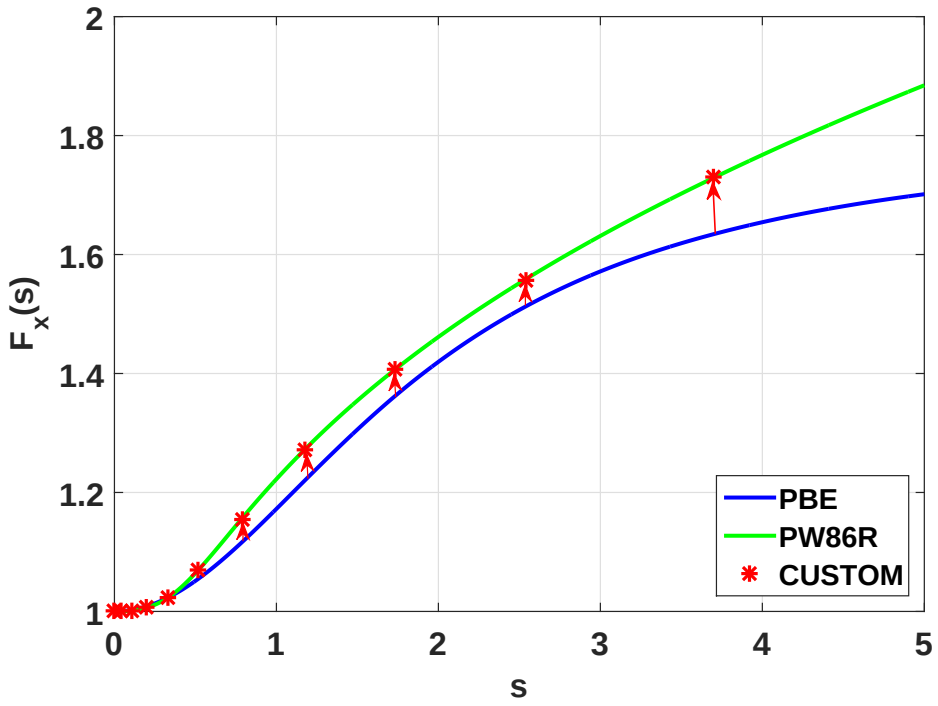


Fig. 4.5 PBE and PW86R exchange enhancement factors (blue and green lines, respectively) along with the GGA exchange enhancement factor values $F_{x\alpha}$ at interpolation points s_α, optimized without constraints to reproduce PW86R exchange energies of 198 water dimers (red asterisks).

water dimer 2-body energies, and 100 water trimer 3-body energies. All reference energies were calculated using the MB-pol potential [4, 6], which is optimized to closely reproduce water energies at the CCSD(T) [109] level of theory.

The monomer geometries used in the optimization of 1-body energies are specified by the two O-H bond lengths r and the angle θ between them. In order to uniformly represent the relevant regions of the configurational space, the points computed correspond to a grid with r= 0.9000, 0.9139, 0.9278, 0.9417, 0.9556, 0.9694, 0.9833, 0.9972, 1.011, and 1.025 Å and θ=97.00°, 98.56°, 100.1°, 101.7°, 103.2°, 104.8°, 106.3°, 107.9°, 109.4°, and 111.0° (Fig. 4.6). In addition, a number of points with a larger and smaller r and a larger and smaller θ were added to characterize other significant regions of the fit.

A set of 330 dimer geometries was chosen from the flexible monomer training set used for the MB-pol 2-body terms [4], which was extracted from a NVT path-integral molecular dynamics (PIMD) simulation of liquid water at ambient temperature and

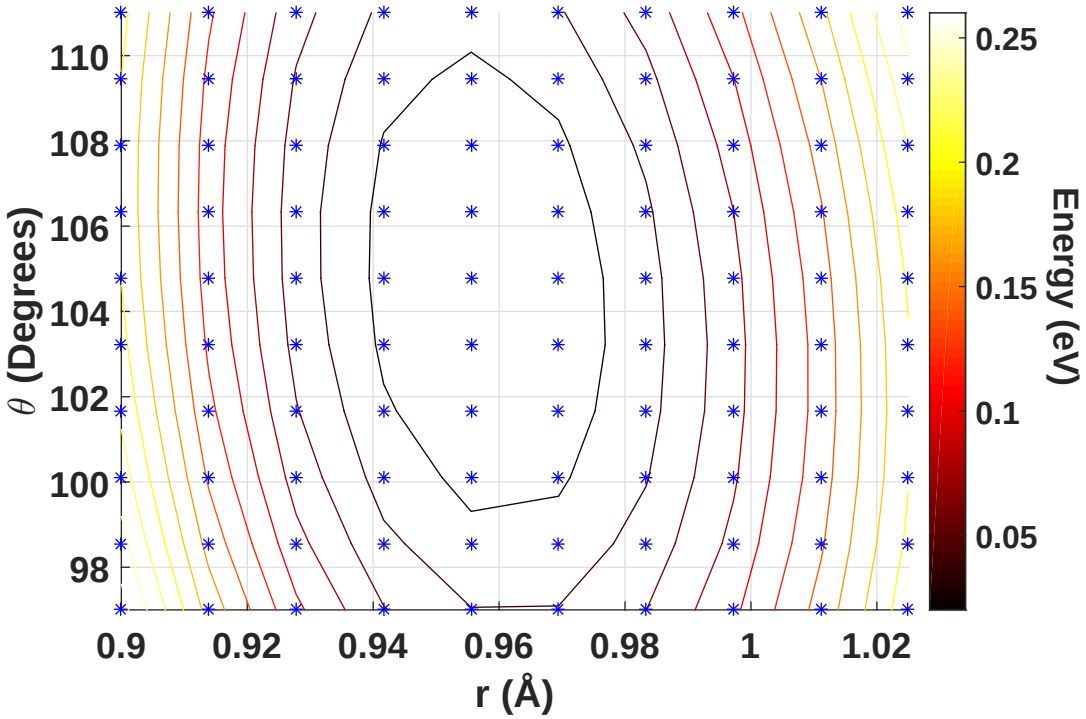


Fig. 4.6 Grid placement for symmetric water monomer geometries and MB-pol energy contours. The y-axis represents the H-O-H angle θ, and the x-axis represents the distance r between O and H. The contour values are at increments of 0.2 eV.

experimental density with the HBB2-pol [5] potential. In order to represent the relevant regions of the configurational space as uniformly as possible, we choose a set that included a variety of combinations of O-O distances and relative monomer orientations (Fig. 4.7).

A set of 100 trimers used in functional optimizations were taken from the short range training set used for the 3-body terms of the MB-pol [6], which were extracted from the following sets: the low-energy subset of the HBB2-pol training set [77]; PIMD simulations of liquid water carried out at ambient conditions with the HBB2-pol potential [5]; PIMD/HBB2-pol simulations of small $(H_2O)_N$ clusters with $N \geq 6$ carried out in the temperature range between 30 and 100 K [7]; and constant pressure-constant temperature (NPT) molecular dynamics simulations of liquid water carried out at ambient conditions with intermediate versions of the MB-pol potential. A set of 100 trimer geometries were chosen such that they make up a distribution of 3-

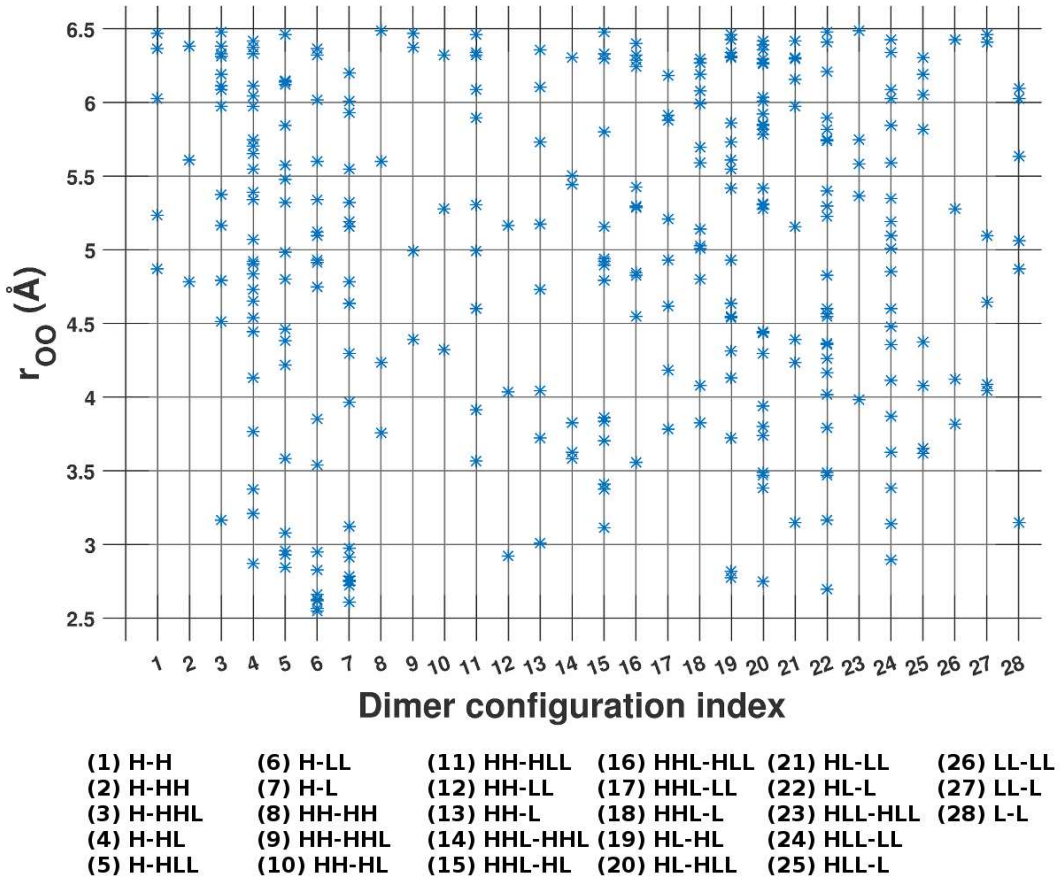


Fig. 4.7 Grid placement for 330 water dimers. The y-axis represents the O-O distance r_{OO}, and the x-axis represents the dimer configuration index. The configurations represented by the indices are shown in the key on the bottom. An "H" represents a hydrogen atom and L represents a lone pair. The letters on each side of the dash represent the portion of the water monomer that is most directly facing the other monomer. For example, in the "H-HLL" configuration, the hydrogen atom (H) of the first monomer is facing the second molecule, while the second monomer has the area between two lone pairs and a hydrogen (HLL) pointed towards the first molecule.

body energies that resembles that of the MB-POL 3-body training set. The range of 3-body energies is shown in Fig. 4.18b of Section 4.2.4.

4.2.4 Optimized functionals

Unconstrained 1-body energy optimization

After testing the DPPS method and choosing a suitable dataset, we first applied it to the optimization of the water monomer energy in 300 geometries, relative to the average energy of the configuration set. Fig. 4.8a shows the $F_x(s)$ optimized for these 1-body (1B) energies, labelled CUSTOM-1B (unconstr.). Also plotted is the $F_x(s)$ of LV-PW86R, the exchange part of the vdw-DF-cx functional, used for the initial functional parameters $F^0_{x\alpha}$ in Eq. 4.18. Note that in Fig. 4.8a, and in all of the following plots of $F_x(s)$, the dots indicate initial and final parameter values, while the lines represent the cubic spline interpolations between the parameter values. For the fit of CUSTOM-1B (unconstr.), Eq. 4.18 was minimized with the error in $F_x(s)$ estimated as $\Delta F = 0.01$ and $\Delta E_n = 3$ meV.

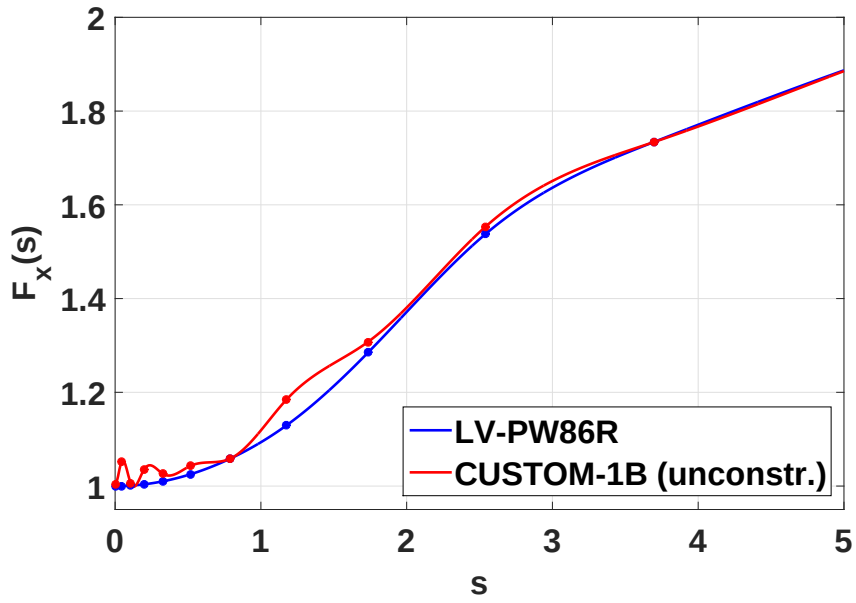
Fig. 4.8b shows the monomer energies predicted by the initial and optimized functionals, plotted against the MB-pol energies. The CUSTOM-1B (unconstr.) functional reduces the root-mean-square (RMS) error of 1-body energies from 12.49 meV/atom to 0.15 meV/atom (See Table 4.1 and Fig. 4.8b). However, despite its optimal performance for 1-body energy calculations, the CUSTOM-1B (unconstr.) $F_x(s)$ has various minima and maxima, making it rather unplausible from a theoretical point of view and unlikely to transfer well to other terms of the total energy and to other systems.

Monomer density analysis in parameter space

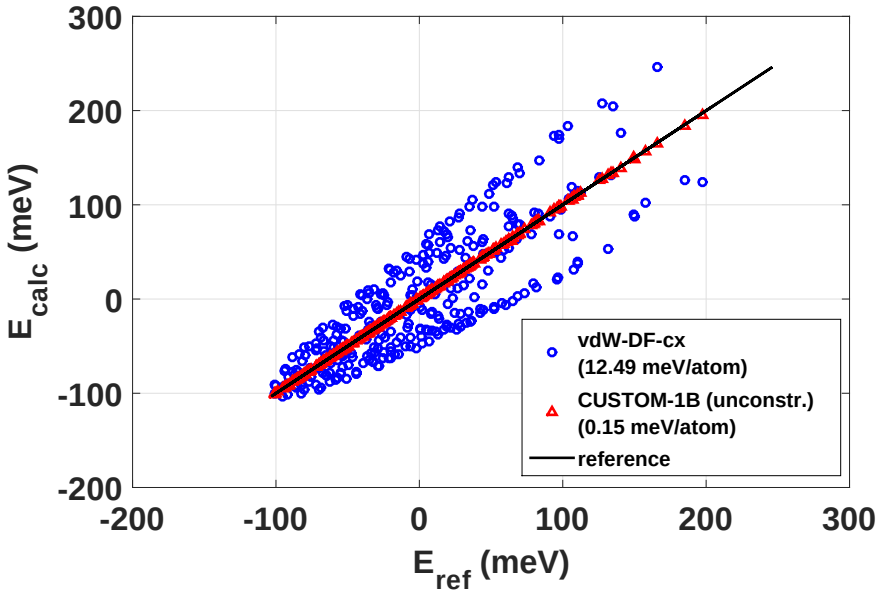
Because the GGA exchange depends on the product of the electron density $\rho(\mathbf{r})$ and the exchange enhancement factor $F_x(s)$, i.e.,

$$E_{xc}[\rho(\mathbf{r})] = \int \rho(\mathbf{r})\ \varepsilon_x^{LDA}(k_F(\mathbf{r}))\ F_x(s)\ d^3\mathbf{r}, \tag{4.19}$$

by studying the changes in $\rho(\mathbf{r})$ upon changes in the monomer geometry, we can get an idea of what adjustments need to be made in $F_x(s)$ to optimize the functional to 1-body energies. It is helpful in this case to transform the electron density from



(a)



(b)

Fig. 4.8 (a) GGA exchange enhancement factor $F_x(s)$, optimized without constraints to reproduce energies calculated with the MB-pol potential [6] of 300 water monomers (CUSTOM-1B (unconstr.), red) and that of vdW-DF-cx (LV-PW86R, blue) and (b) energies of 300 monomers calculated with the CUSTOM-1B functional (red) and with the vdW-DF1-cx functional (blue) against the MB-pol reference energies. Dots indicate parameter values, and lines represent the cubic interpolations between the points. The RMS errors are shown in the legends.

real space ($\rho(\mathbf{r})$) to the space determined by k_F and k_G ($\rho(k_F, k_G)$), where the GGA exchange energy densities $\varepsilon_x(k_F, k_G)$ can be defined, and to "parameter space" or s space, where $F_x(s)$ is defined. The electron density $\rho(k_G, k_F)$ is calculated by integrating over real space:

$$\rho(k_F, k_G) = \int d^3\mathbf{r}\rho(\mathbf{r})\delta(k_F - (3\pi^2\rho(\mathbf{r}))^{1/3})\delta(k_G - |\nabla\rho(\mathbf{r})|/\rho(\mathbf{r})) \tag{4.20}$$

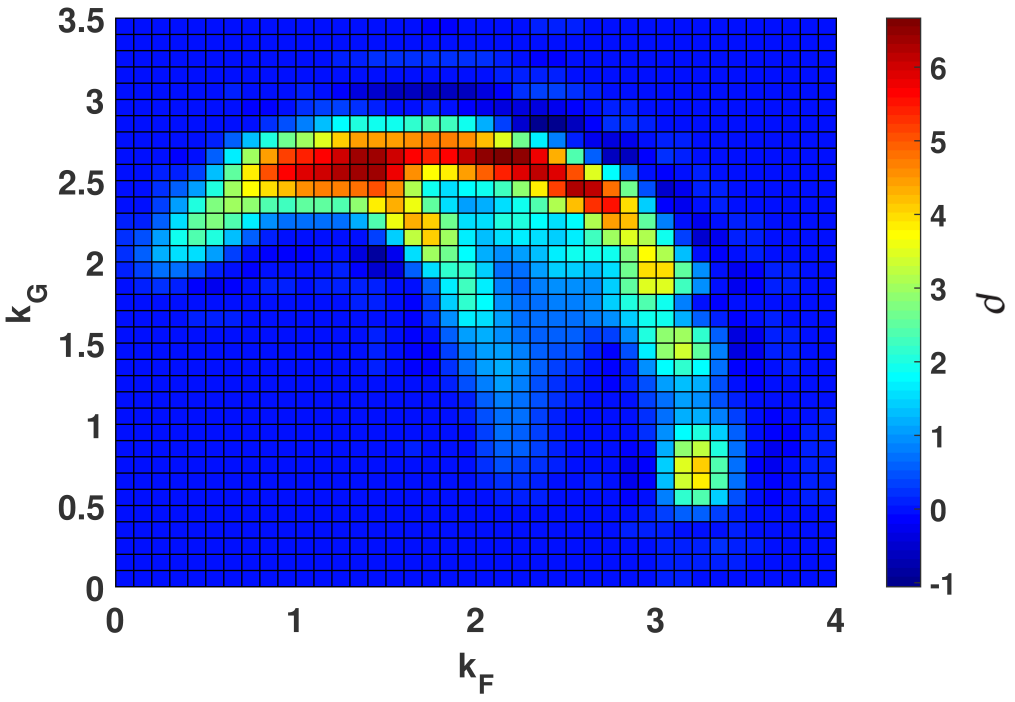
and

$$\rho(s) = \int\int dk_F dk_G \rho(k_F, k_G)\delta(s - k_G/2k_F). \tag{4.21}$$

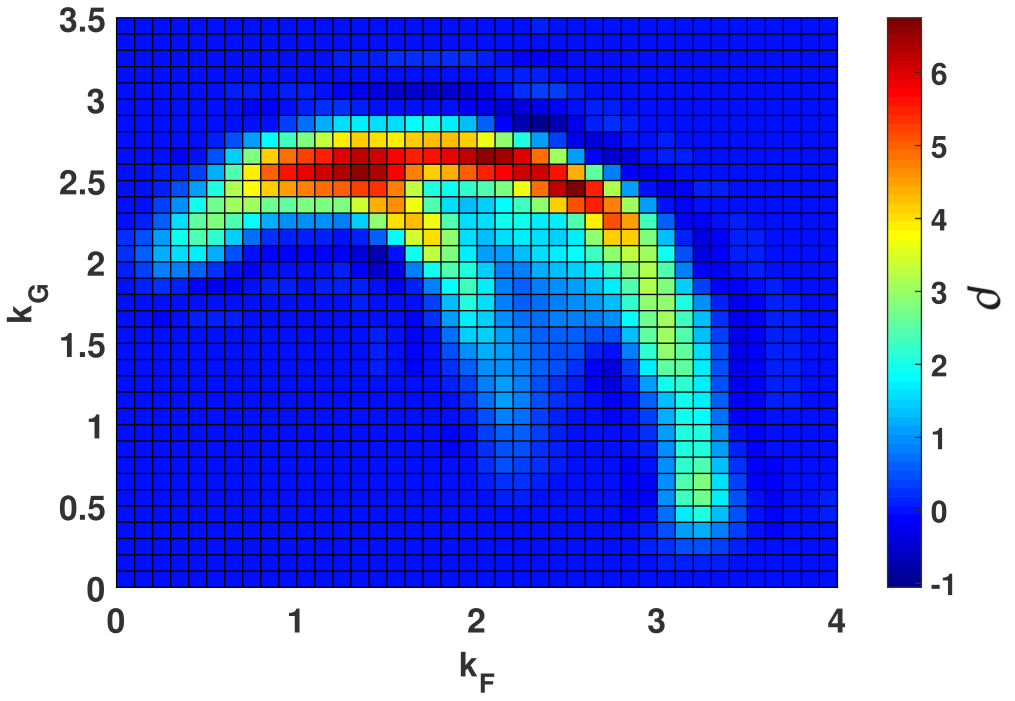
Before looking at density changes, we first look at the density of a water monomer in equilibrium. A plot of $\rho(k_F, k_G)$ for a monomer in equilibrium according to the MB-pol potential, with 300 Ry and 600 Ry mesh cutoffs, is shown in Fig. 4.9. With the smaller mesh cutoff, the plot of $\rho(k_F, k_G)$ can be misleading, especially where there are smaller, concentrated density values. This is seen mostly in the bottom tails of the plots in Fig. 4.9. When evaluated with a 300 Ry mesh cutoff, the right tail of $\rho(k_F, k_G)$, which corresponds to the density of the core electrons, the density is collected into more concentrated areas, indicating non-continuity in the density gradient. The plot of $\rho(k_F, k_G)$ evaluated with a 600 Ry mesh cutoff, however, shows a more continuous tail, which is more realistic.

Next we look at changes in the electron density and how the mesh cutoff affects the form of $F_x(s)$. Fig. 4.10 and Fig. 4.11 show the changes in density $\Delta\rho(k_F, k_G)$ and $\Delta\rho(s)$, respectively, upon changing from the equilibrium geometry of the vdw-DF-cx functional to that of MB-pol, with 300 Ry and 600 Ry mesh cutoffs. In the MB-pol equilibrium geometry, the O-H bonds are shorter (see Table 4.2). This causes the electron density to be closer to the oxygen atom, with values that increase more rapidly in real space, resulting in higher electron density gradients (and therefore higher values of k_G). This can be seen in Fig. 4.10, where the changes indicate density values that shift from lower to higher values of k_G.

With a 300 Ry mesh cutoff, some artificial changes appear in the density, especially around lower values of s, where the changes in the density come from changes in the core electron density. These changes are unnaturally large with the lower mesh cutoff and correspond to changes which are negligible in reality. This is seen more clearly in Fig. 4.11, which shows plots of the changes in $\rho(s)$. The oscillations with a 600 Ry mesh cutoff in Fig. 4.11b are smoothed out and less pronounced in comparison to the plots done with a 300 Ry mesh cutoff in Fig. 4.11a. Because a low mesh cutoff leads to inaccurate changes in $\rho(s)$, it will also be prob-

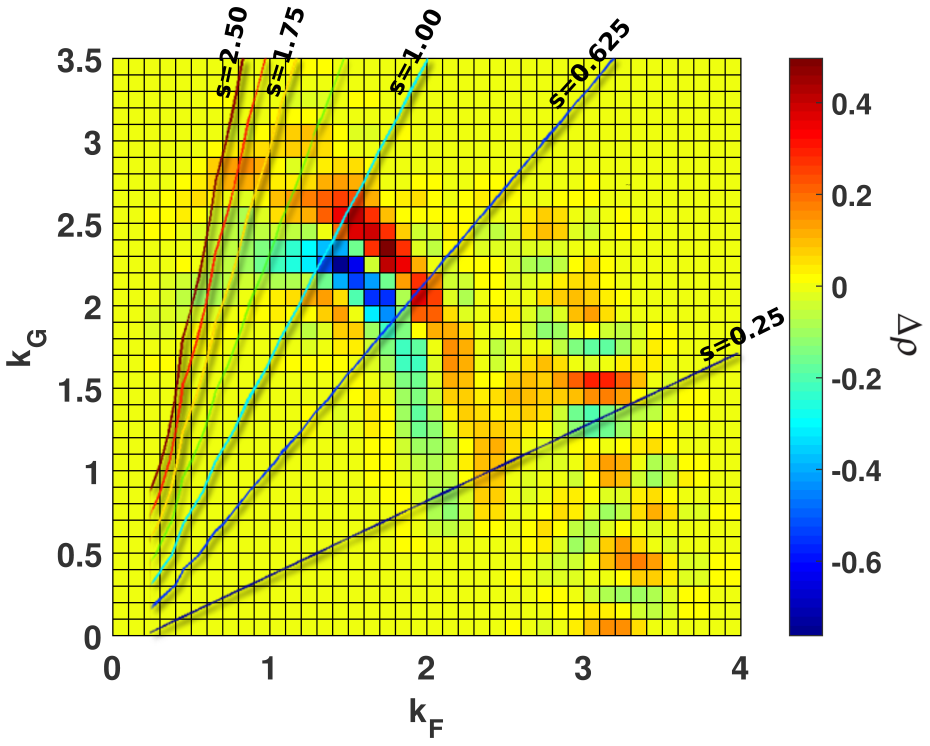


(a)

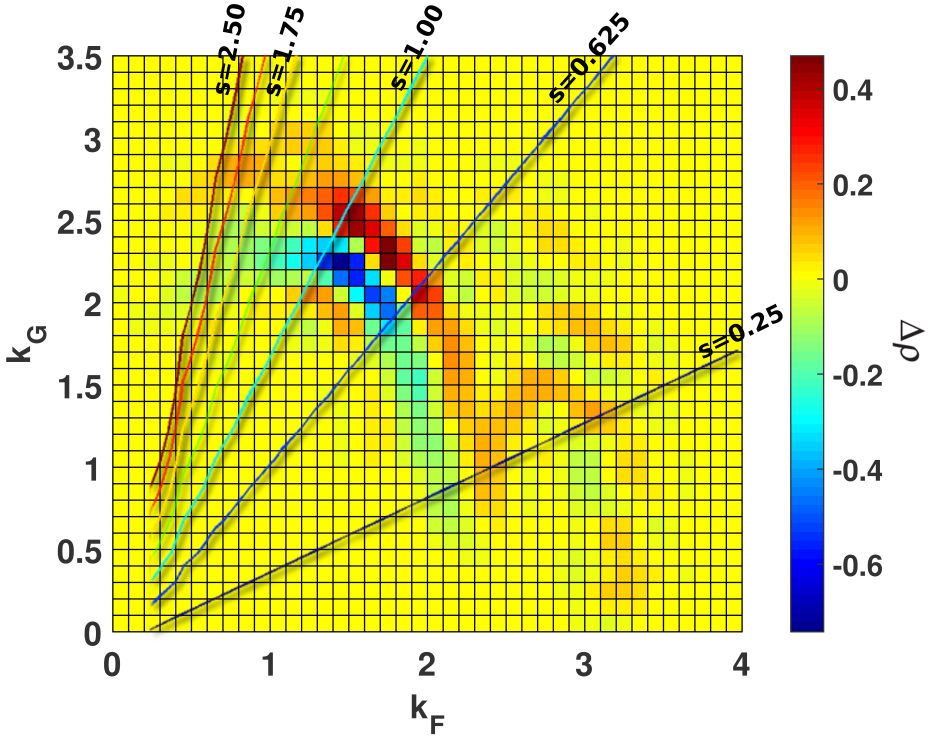


(b)

Fig. 4.9 The electron density $\rho(k_F, k_G)$ (Eq. 4.20) for a monomer at its equilibrium geometry, with a 300 Ry mesh cutoff (a) and a 600 Ry mesh cutoff (b).

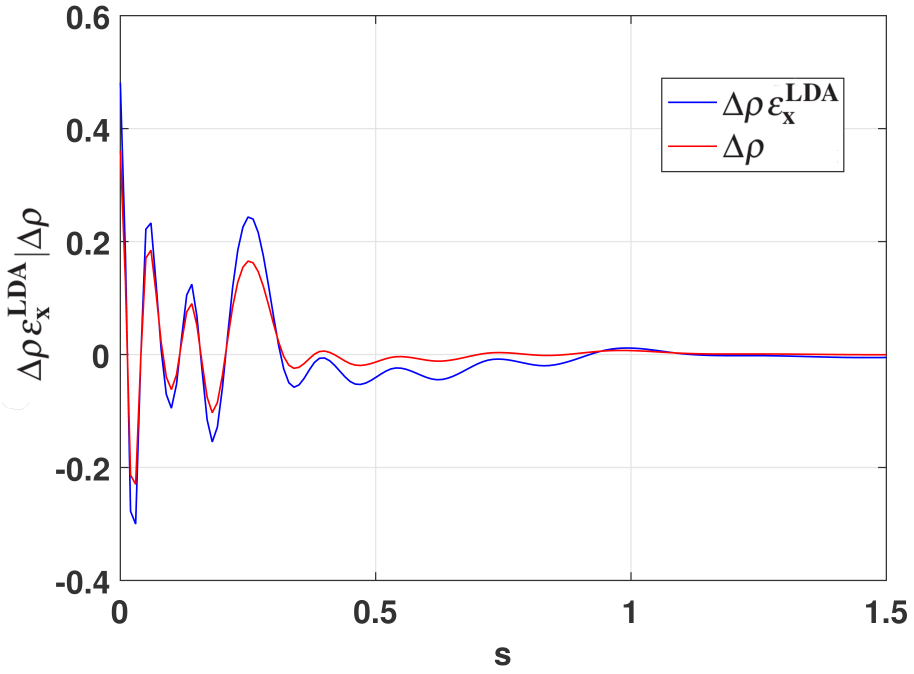


(a)

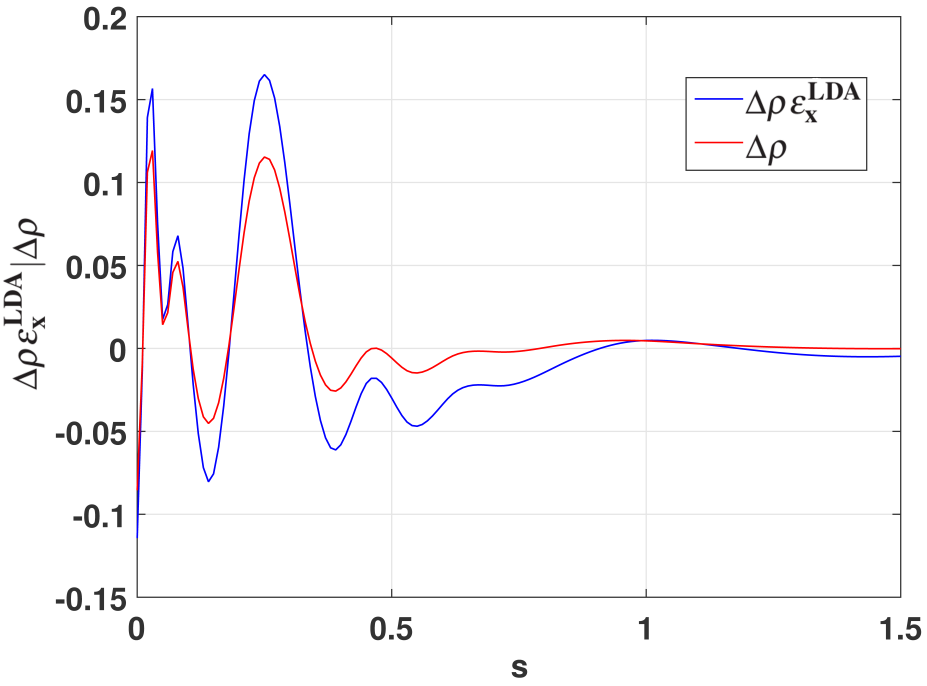


(b)

Fig. 4.10 Changes in the electron density $\rho(k_F, k_G)$ upon moving from the monomer equilibrium geometry of vdw-DF-cx functional to that of MB-pol, with a 300 Ry mesh cutoff (a) and a 600 Ry mesh cutoff (b). The lines represent constant values of s. The equilibrium geometries can be found in Table 4.2.



(a)



(b)

Fig. 4.11 Changes in the electron density $\rho(s)$ (Eq. 4.21) upon changing from the water monomer equilibrium geometry of vdw-DF-cx functional to that of MB-pol, with (a) 300 Ry and (b) 600 Ry mesh cutoffs. Also represented are the changes in the electron density $\rho(s)$ multiplied by the LDA exchange energy density ε_x^{LDA} to show the weight of the changes.

lematic when looking for changes in $F_x(s)$. Since many of the large changes in density $\rho(k_F, k_G)$ upon shortening the covalent bonds for low values of s with a 300 Ry mesh cutoff are artificial, the oscillations in the $F_x(s)$ optimized with a 300 Ry mesh cutoff in Fig. 4.13a are partially an artifact of the low mesh cutoff and do not represent true changes that are needed in the functional. However, the use of a 600 Ry mesh cutoff for energy calculations would add a large computational cost. A way to avoid artificial changes in $F_x(s)$ brought upon by an insufficient mesh cutoff, without adding additional computational cost, is to optimize with Bayesian constraints, as described in the following subsection.

A critical factor to consider in the analysis of the monomer density is the ability of the density changes to transfer to parameter changes. Fig. 4.10b shows that most changes in $\rho(k_F, k_G)$ occur along a constant value of s. However, these changes cannot be "noticed" by a functional that depends only on s. The small density differences that change the value of s will result in changes in $F_x(s)$ that overcompensate for this, adding to the artificiality of the new functional. Fig. 4.11 clearly shows that some of the largest changes in s are around $s \sim 0.25$. However, the plot in Fig. 4.10b shows that the changes around these values of s are in reality among the smallest changes needed for reaching the true equilibrium geometry. In fact, the most significant changes, which occur at higher values of s, do not appear in Fig. 4.11, because they occur along lines of constant s. This conclusion, based on the density changes between the DFT predicted and reference monomer equilibrium geometry, tells us that it will be especially difficult to optimize a functional for 1-body energies simultaneously with another type of dataset (2- and 3-body terms). The large, unnatural changes brought about by the optimization to 1-body terms will compete with the optimization of the other datasets.

Studying these changes in electron density also exposes the physical root of a fundamental problem with non-hybrid GGA functionals: their inability to correct self-interaction errors. Self-interaction errors, as explained in Section. 3.2.2, are contained in the 1-body terms and therefore would be fixed upon the correction of these terms. However, a functional that depends only on s cannot produce the necessary changes in the 1-body energies, due to the nature of the electron density.

Constrained 1-body energy optimization

In order to obtain a more realistic functional, we included Bayesian constraints in another optimization to 1-body energies. In order to apply these constraints, Eq. 4.10 in Section 4.1.3 is maximized (unlike the unconstrained case, in which Eq. 4.2 is minimized). In the case of optimization of only GGA exchange, ε_α and ε_α^0 in Eq. 4.12 can be replaced with $F_{x\alpha}^0$ and $F_{x\alpha}^0(s)$, respectively, so that it becomes

$$\mathscr{P}(theory) = C_2 \exp[-\sum_{\alpha\beta} \frac{1}{2}(F_{x\alpha} - F_{x\alpha}^0) Cov_{\alpha\beta}^{-1} (F_{x\beta} - F_{x\beta}^0)], \tag{4.22}$$

where the indices α denote interpolation points s_α. Then ε_α^i is replaced by $F_{x\alpha}^i(s)$ in Eq. 4.13 so that it becomes

$$F_{x\alpha}^0 = \sum_i w_i F_{x\alpha}^i, \tag{4.23}$$

where the values $F_{x\alpha}^i$ are determined by the different forms of $F_x(s)$ used in pre-existing *ab initio* GGA exchange functionals. The covariance matrix becomes

$$Cov_{\alpha\beta} = \sum_i w_i (F_{x\alpha}^i - F_{x\alpha}^0)(F_{x\beta}^i - F_{x\beta}^0). \tag{4.24}$$

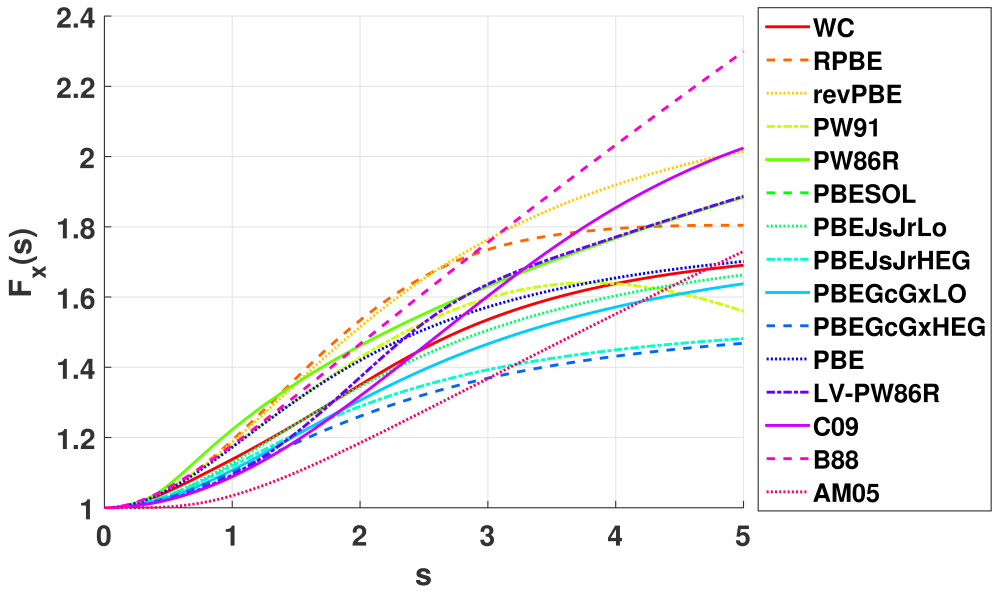
For all constrained optimizations discussed in this chapter, $F_{x\alpha}^i$ were taken from the following functionals (with equal weights):

- PW91 [104]
- PBE [98]
- RPBE [38]
- revPBE [135]
- WC [132]: the Wu-Cohen modification of PBE functional
- PBESOL [100]
- AM05 [74]
- PBEJsJrLO [95, 88]: the PBE functional with parameters β, μ, and κ fixed by the jellium surface (Js), jellium response (Jr), and Lieb-Oxford bound (LO) criteria

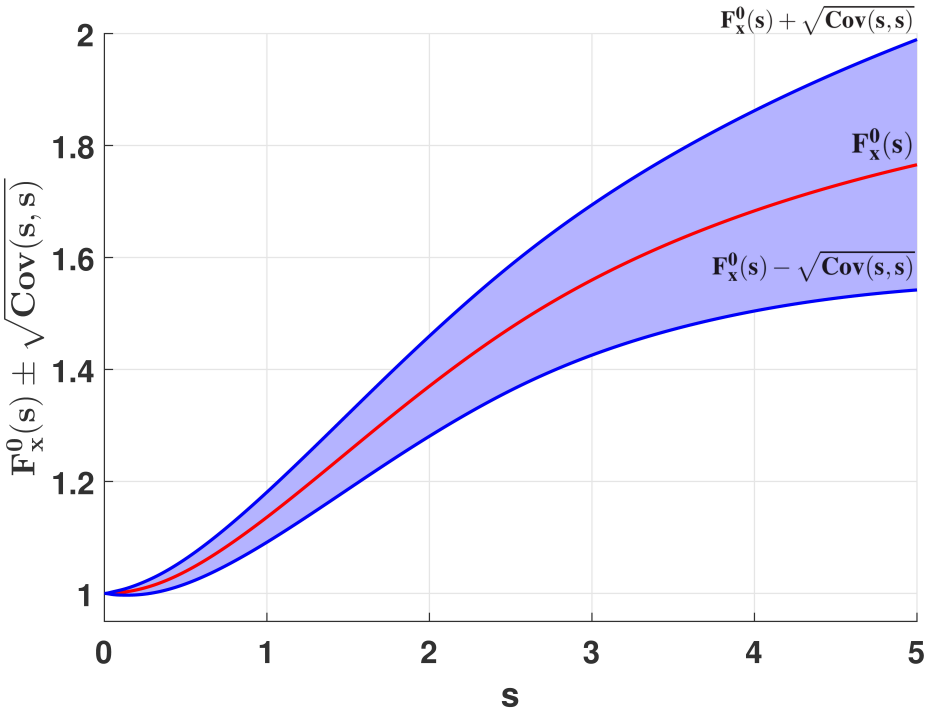
- PBEJsJrHEG: the same as PBEJsJrLO with parameter κ fixed by the Lieb-Oxford bound for the low density limit of the homogeneous electron gas (HEG)
- PBEGcGxLO: the same as PBEJsJrLO with parameters β and μ fixed by the gradient expansion of correlation (Gc) and exchange (Gx), respectively
- PBEGcGxHEG: the same as PBEJsJrLO with parameters β, μ, and κ fixed by the Gc, Gx, and HEG criteria, respectively
- PW86R [102, 85]
- B88 [10]: the exchange functional of optB88-vdW, which is the modified version B88 exchange by Klimes, Bowler, and Michaelides [52]
- C09 [19]
- LV-PW86R [14]: the exchange of the vdW-DF-cx functional.

Fig. 4.12a shows the form of $F_x(s)$ used in each one of the functionals listed here. The average of these enhancement factors $F_x^0(s) \pm$ the diagonal of the covariance matrix $Cov(s,s')$ is plotted in Fig. 4.12b to show the deviation in the forms of $F_x^i(s)$ from the average. A contour plot of the whole covariance matrix $Cov(s,s')$ is given in Fig 4.12c.

Note that, as shown in Fig. 4.12b, the values of $F_x^i(s)$ deviate very little from $F_x^0(s)$ for small values of s. This means that the constraints on $F_x(s)$ will be stricter in this region. For values of $s = 0$, every form of $F_x^i(s)$ is equal to one, indicating that each form obeys the HEG limit and constraints are very strict at this point. However, for larger values of s, there is a larger deviation from $F_x^0(s)$, and consequently, a constrained functional will be allowed much more flexibility in this region. Fig. 4.13 shows $F_x(s)$ optimized for water monomers only, this time implementing Bayesian constraints (CUSTOM-1B), along with the LV-PW86R enhancement factor. Unlike the unconstrained case, the CUSTOM-1B $F_x(s)$ is a monotonically increasing function. However, by including the constraints, we obtain a larger error in the energy than with the unconstrained fit, with an RMS error of 0.59 meV/monomer (See Table 4.1 and Fig. 4.13b).

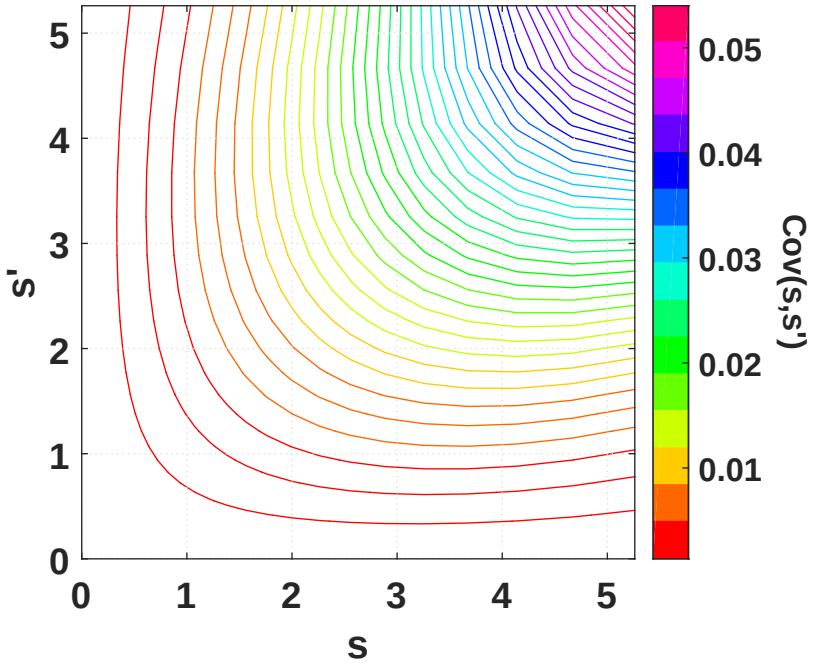

2.4
2.2
2
1.8
1.6
1.4
1.2
1
0
1
2
3
4
5
s
$F_x(s)$
WC
RPBE
revPBE
PW91
PW86R
PBESOL
PBEJsJrLo
PBEJsJrHEG
PBEGcGxLO
PBEGcGxHEG
PBE
LV-PW86R
C09
B88
AM05

(a)


$F_x^0(s) + \sqrt{Cov(s,s)}$
$F_x^0(s)$
$F_x^0(s) - \sqrt{Cov(s,s)}$
$F_x^0(s) \pm \sqrt{Cov(s,s)}$
2
1.8
1.6
1.4
1.2
1
0
1
2
3
4
5
s

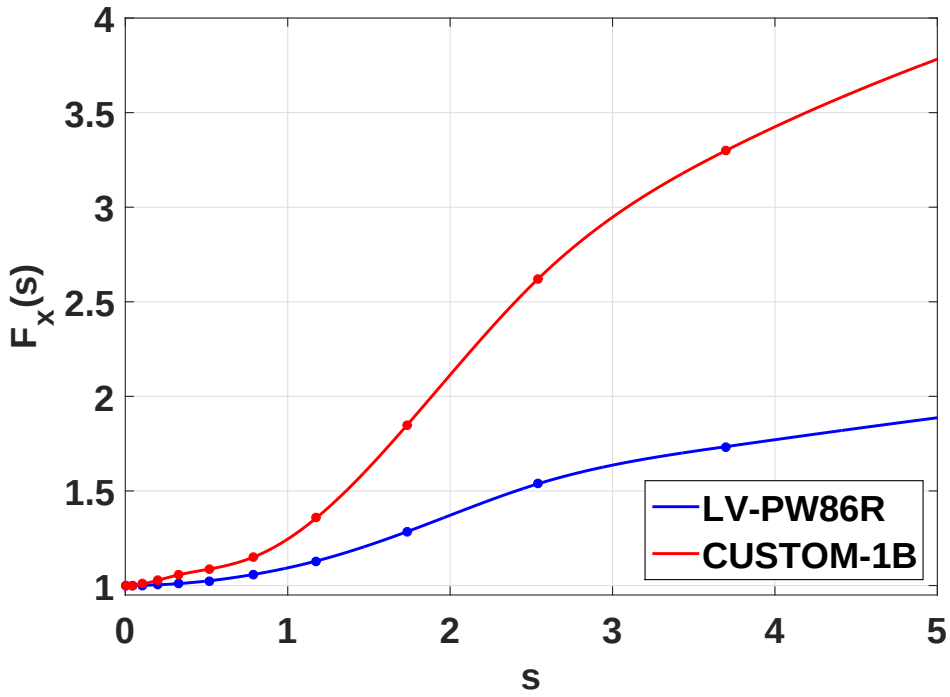
(b)

Fig. 4.12

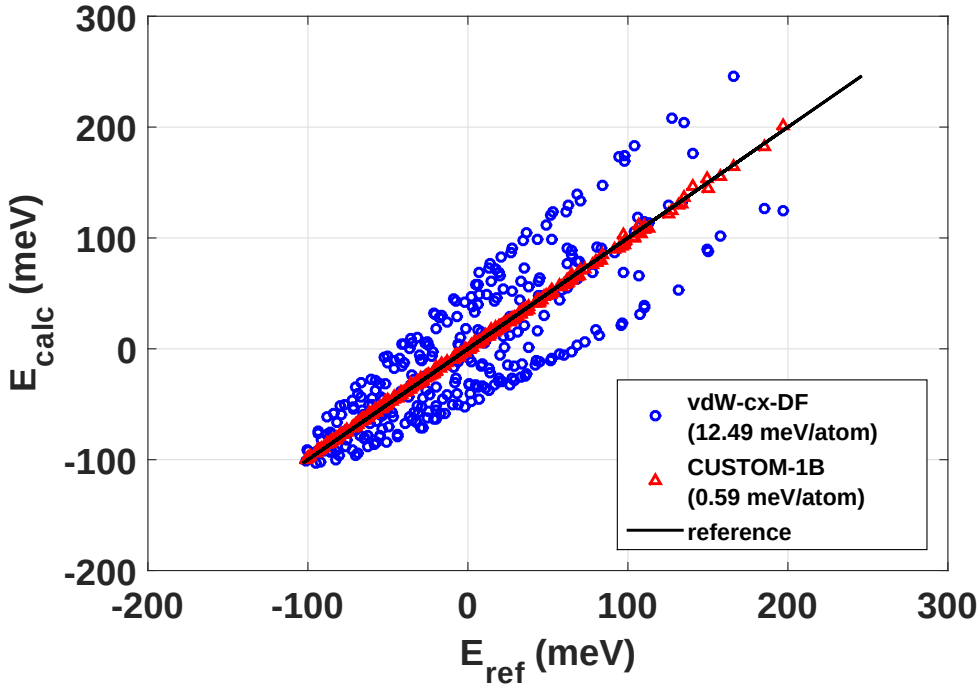


(c)

Fig. 4.12 (a) Plot of all forms of $F_x(s)$ used to calculate the average $F_x^0(s)$ (Eq. 4.12b) and the covariance matrix $Cov(s,s')$ (Eq. 4.24) in all Bayesian constrained DPPS optimizations presented in this thesis and (b) plot of $F_x^0(s) \pm$ the diagonal of $Cov(s,s')$, which is plotted in (c). The contour values in (c) are at increments of 1.35×10^{-3}.

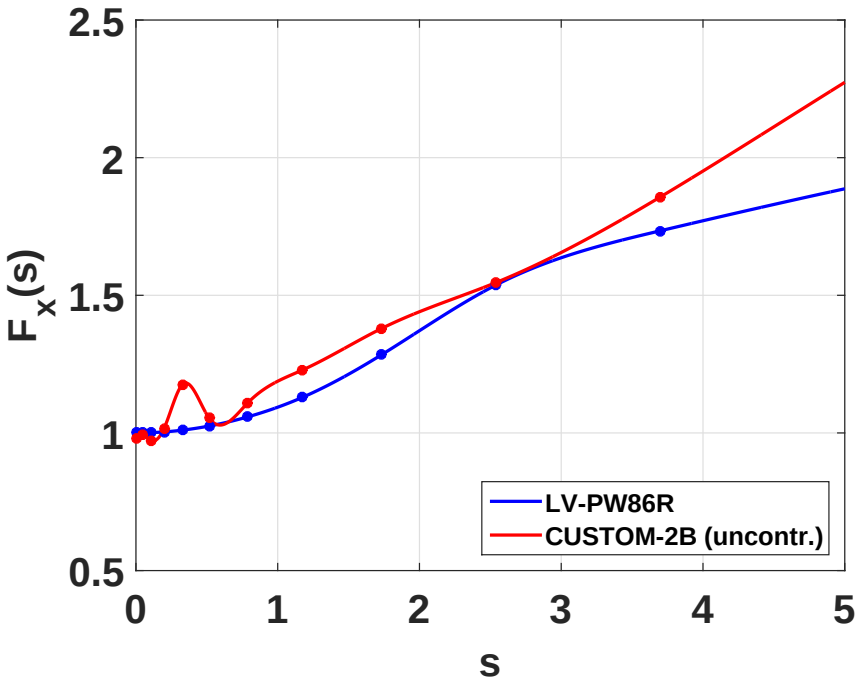


(a)

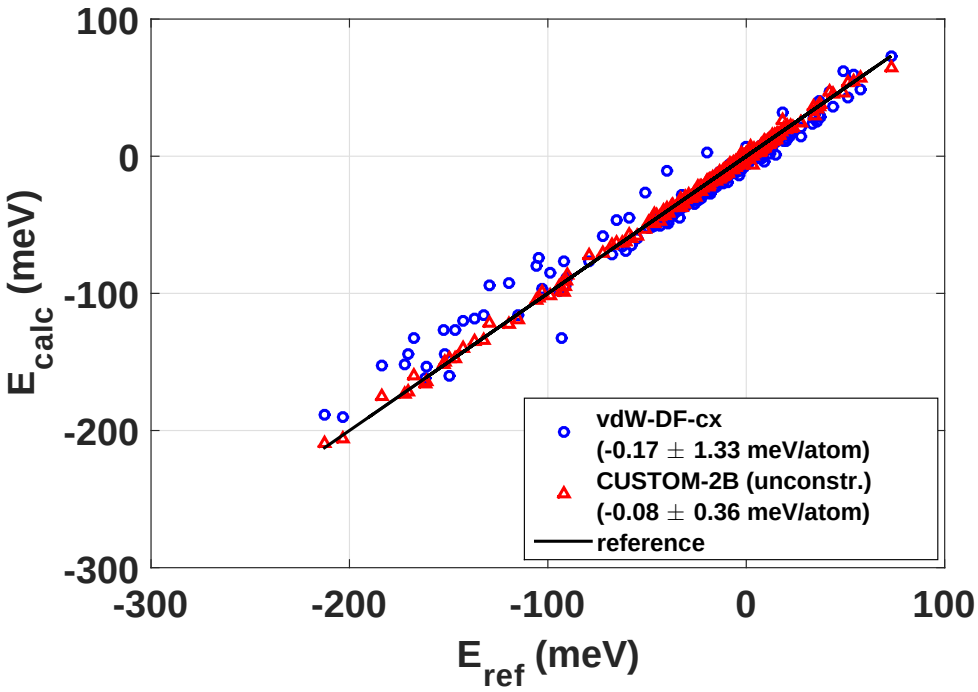


(b)

Fig. 4.13 The plots in (a) and (b) are the same as those in Figures 4.13a and 4.8b, with the exception that Bayesian constraints were used to fit the functional CUSTOM-1B.



(a)



(b)

Fig. 4.14 (a) GGA exchange enhancement factor $F_x(s)$, optimized without constraints to reproduce 2-body energies calculated with the MB-pol potential [6] of 330 dimers (CUSTOM-2B (unconstr.), red) and that of vdW-DF-cx (LV-PW86R, blue) and (b) 2-body energies of the same 330 dimers calculated with the CUSTOM-2B functional (red) and with the vdW-DF1-cx functional (blue) plotte against the MB-pol reference energies. The average errors ± RMS errors are shown in the legends.

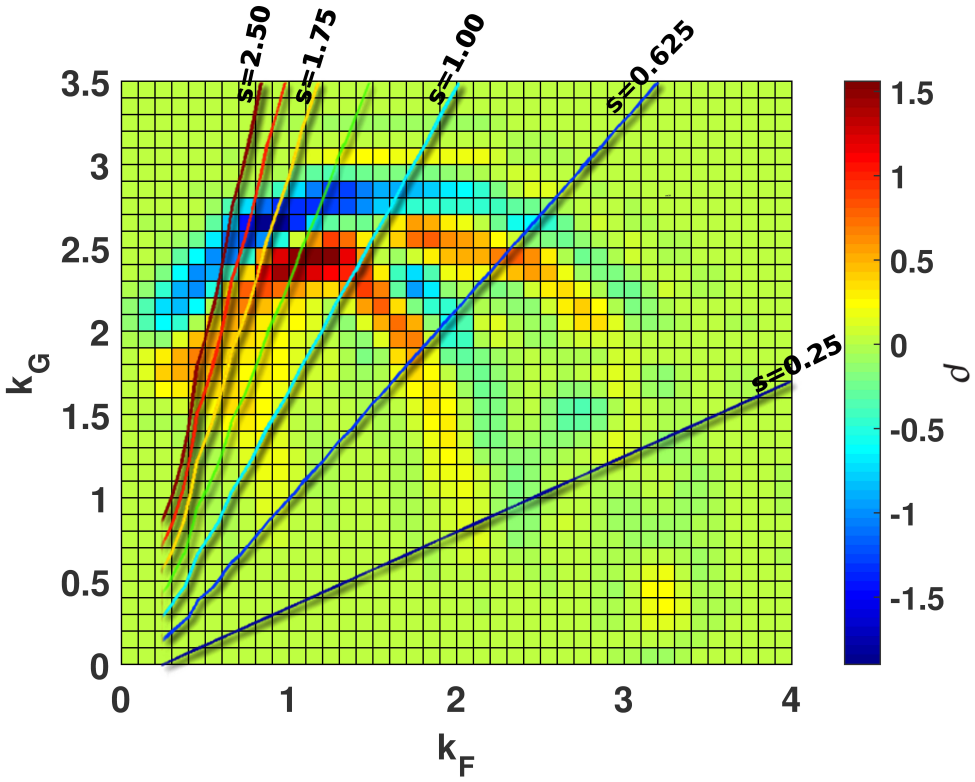
Unconstrained 2-body energy optimization

Next, we optimized $F_x(s)$ to reproduce the reference set of 2-body energies of 330 water dimers. The form of the GGA exchange was optimized by minimizing Eq. 4.18, with $\Delta F = 0.01$ and $\Delta E_n = 0.3$ meV. The optimized functional is labelled CUSTOM-2B (unconstr.), where 2B stands for "2-body". The CUSTOM-2B (unconstr.) and LV-PW86R forms of $F_x(s)$ are plotted together in Fig. 4.14a. As in the case of CUSTOM-1B (unconstr.), the optimized $F_x(s)$ contains oscillations. In addition, it does not obey the basic constraint that should be obeyed by all GGA functionals according to the HEG limit, which requires that $F_x(0) = 1$. However, it reduces the error from -0.17 $\pm$ 1.32 meV/atom to -0.08 $\pm$ 0.36 meV/atom (Fig. 4.14b). Most of the remaining error cannot be reduced by changing the dependence on s only, as discussed in the case of the 1-body energies. Although in the case of the dimer, this problem is not as extreme, as seen in the next subsection in which the analysis of the dimer density is discussed. In other words, the remaining error is intrinsic to the GGA exchange functional.

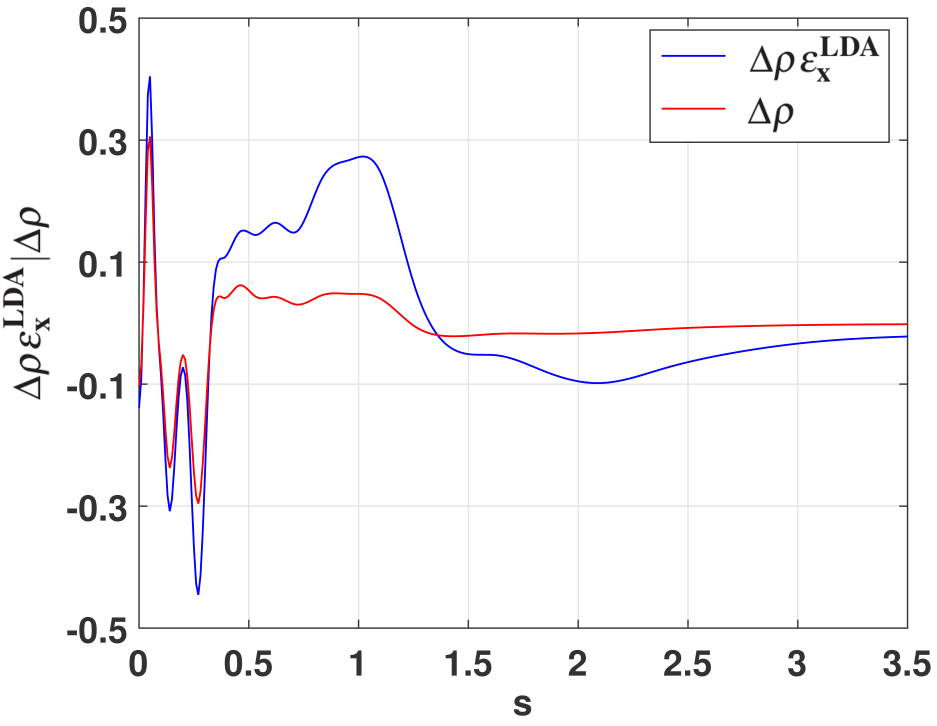
Dimer density analysis

An analysis of the changes in electron density between a dimer and its isolated monomers tells us the important regions for changes in $F_x(s)$ in a 2-body energy optimization. The changes in $\rho(k_F, k_G)$ and $\rho(s)$ upon the formation of a dimer with a mesh cutoff of 600 Ry are shown in Fig. 4.15. Here we see the same problem that occurred in the case of the monomer; there are changes in the density that cannot be reflected in changes in s. However, in the case of dimer formation, there are also significant changes in the electron density along changing values of s. It can be seen that some of the significant changes in $\rho(s)$ (Fig. 4.15b) correspond to the changes seen in $\rho(k_F, k_G)$ (Fig. 4.15a) around $s = 1$. This suggests that the changes in the optimized $F_x(s)$ should better reflect the required physical changes needed to get closer to the reference than in the case of the monomers.

There are significant changes in the shape of $F_x(s)$ for $0 < s < 3$, as expected based on the electron density changes upon dimer formation in Fig. 4.15. The density in Fig. 4.15b suggests that most changes in $F_x(s)$ to reproduce reference 2-body energies will be for $s < 1.5$, but there should also be some smaller changes for $1.5 < s < 3$.

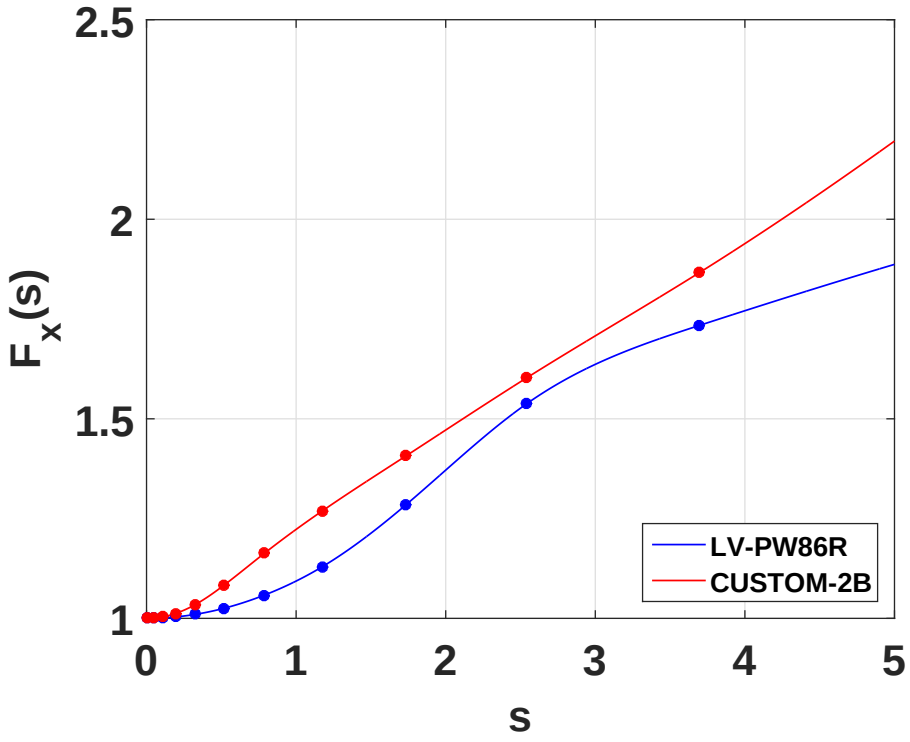


(a)

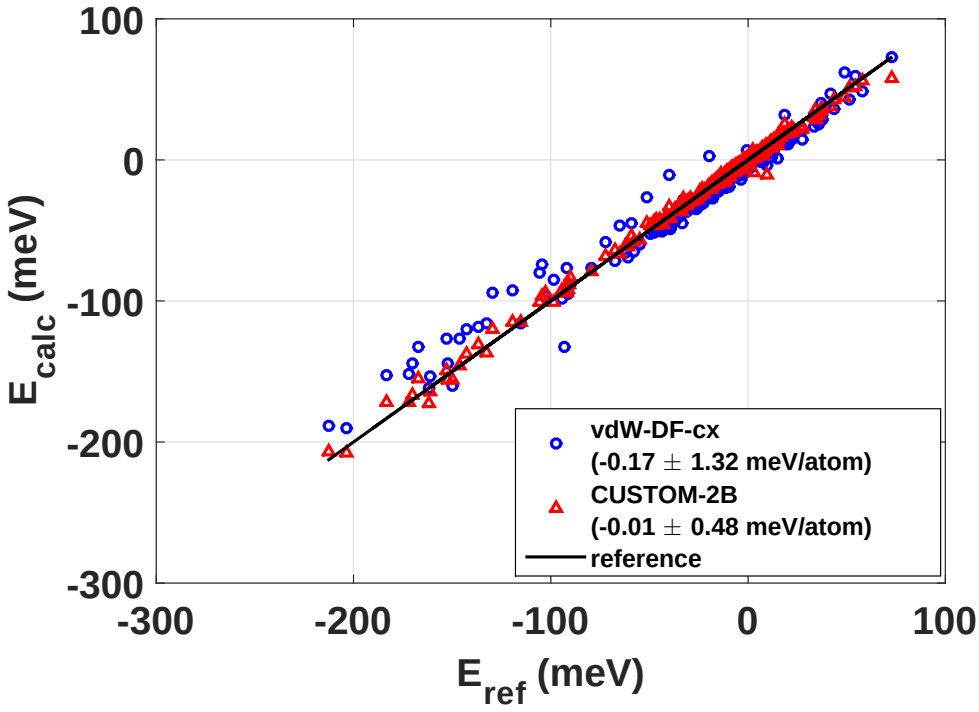


(b)

Fig. 4.15 Changes in the electron density $\rho(k_F, k_G)$ (a) and $\rho(s)$ (b) upon the formation of a dimer at the equilibrium geometry, calculated with a 600 Ry mesh cutoff. The lines represent constant values of s.



(a)



(b)

Fig. 4.16 The plots in (a) and (b) are the same as those in Figures 4.14a and 4.14b, with the exception that Bayesian constraints were used to fit the functional CUSTOM-2B.

Constrained 2-body energy optimization

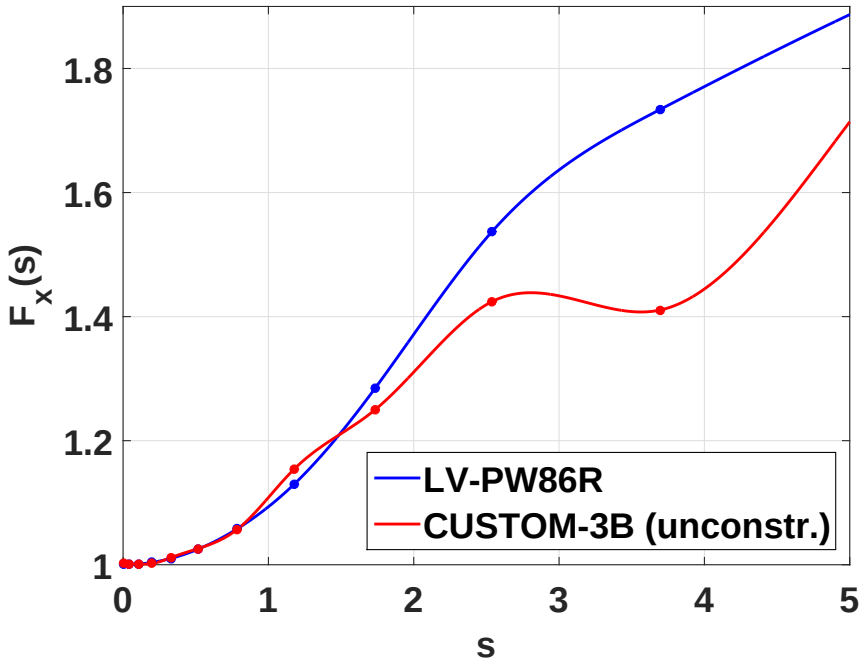
Fig. 4.16a shows $F_x(s)$ optimized for the set of 2-body energies of 330 dimers, using Bayesian constraints (CUSTOM-2B), along with the LV-PW86R enhancement factor. In the constrained case, Eq. 4.10 was maximized with the parameters $F_{x\alpha}$ and $\Delta E_n = 0.3$ meV. The CUSTOM-2B $F_x(s)$ satisfies the basic constraints that were not satisfied in the unconstrained case. However, by including the constraints, we obtain a larger error in the energy than with the unconstrained fit. With CUSTOM-2B, the average error of the dimer 2-body energies is reduced to -0.01 $\pm$ 0.48 meV/atom (See Fig. 4.16b and Table 4.1).

3-body energy optimization (constrained and unconstrained)

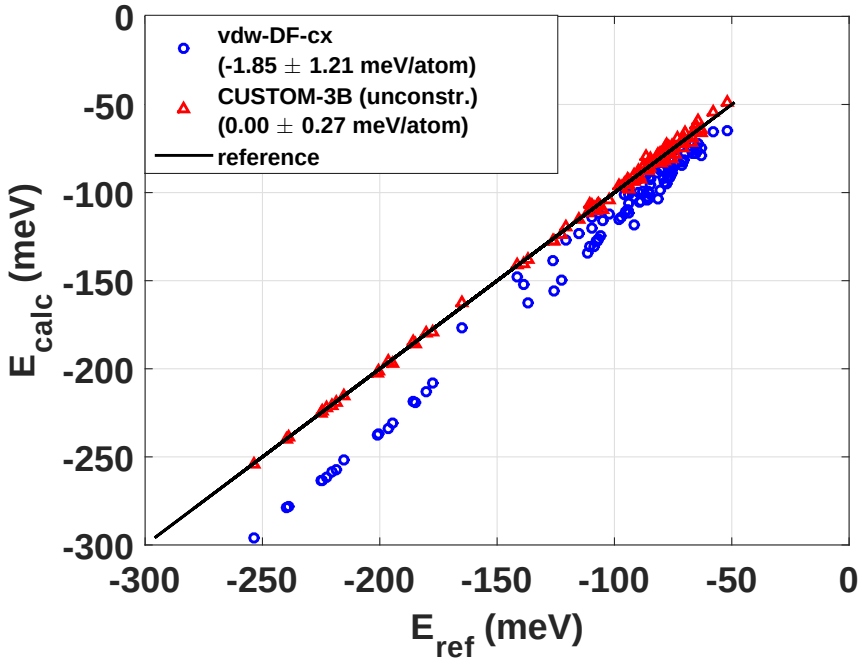
Next, $F_x(s)$ was optimized, without constraints, to a set of 3-body energies of 100 water trimers. The optimized functional is labelled CUSTOM-3B, where 3B represents "3-body". The form of the GGA exchange was optimized by minimizing Eq. 4.18, with $\Delta F = 0.01$ and $\Delta E_n = 0.5$ meV. The CUSTOM-3B form of $F_x(s)$ is plotted in Fig. 4.17a along with the LV-PW86R enhancement factor. Once again, the optimized functional clearly does not obey physical constraints. However, with the optimized functional, the RMS error of the 3-body energies is reduced from -1.85 $\pm$ 1.21 meV/atom to 0.00 $\pm$ 0.27 meV/atom (See Fig. 4.17b and Table 4.1). As seen in the case of the 1- and 2-body energies, there is a remaining error due to the required changes that cannot be made with a functional that depends only on s. With the inclusion of Bayesian constraints in the optimization, by maximizing Eq. 4.10 for $F_{x\alpha}$ with $\Delta E_n = 0.5$ meV, the large valley in the $F_x(s)$ of CUSTOM-3B (unconstr.) is corrected so that the functional is monotonically increasing (Fig. 4.18a). The constrained functional, labelled CUSTOM-3B, reduces the RMS error of the 3-body energies to 0.00 $\pm$ 0.31 meV/atom (See Fig. 4.18b and Table 4.1).

Simultaneous 2-body and 3-body energy optimization (constrained and unconstrained)

To study the changes needed in $F_x(s)$ to correct 2- and 3-body terms simultaneously, we optimized it for a dataset of 2-body energies of 330 dimers and 3-body energies of 100 trimers calculated with the MB-pol potential, without constraints. This was done by minimizing Eq. 4.18 with $\Delta F = 0.01$, $\Delta E_n = 0.3$ meV for the set of 2-body

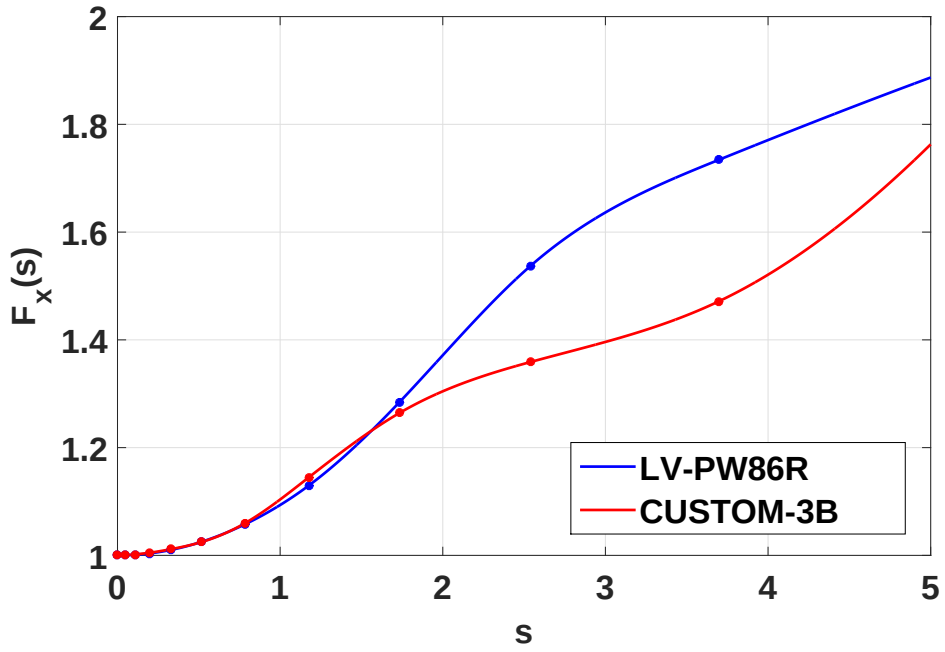


(a)

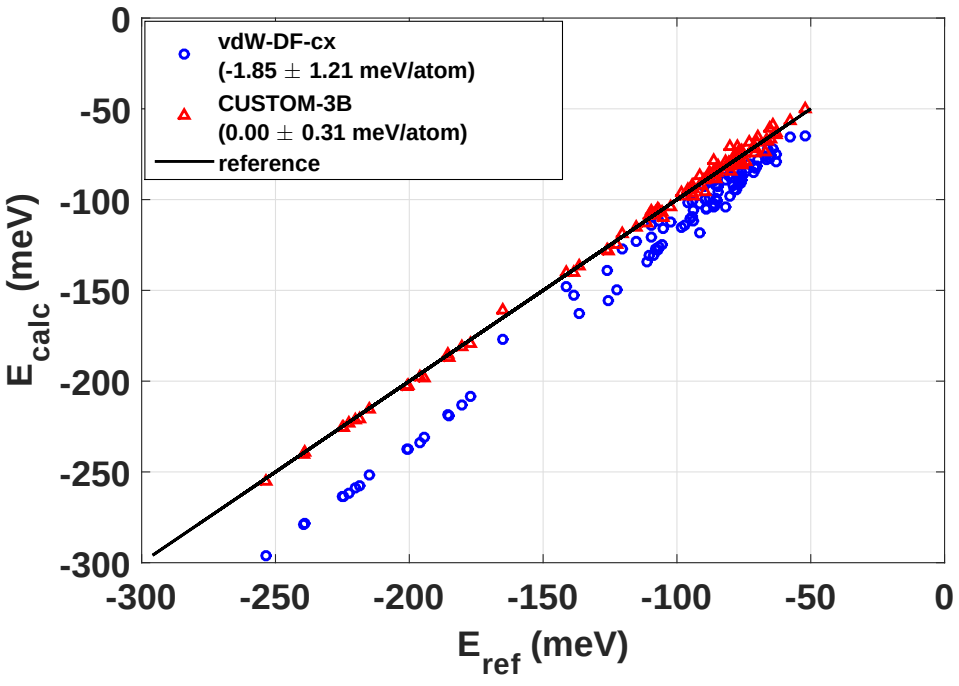


(b)

Fig. 4.17 (a) GGA exchange enhancement factor $F_x(s)$, optimized without constraints to reproduce MB-pol [6] 3-body energiesl of 100 water trimers (CUSTOM-3B (unconstr.), red) and that of vdW-DF-cx (LV-PW86R, blue) and (b) 3-body energies of 100 trimers calculated with the CUSTOM-1B functional (red) and with the vdW-DF1-cx functional (blue) plotted against the MB-pol reference energies. The average errors ± RMS errors are shown in the legends.

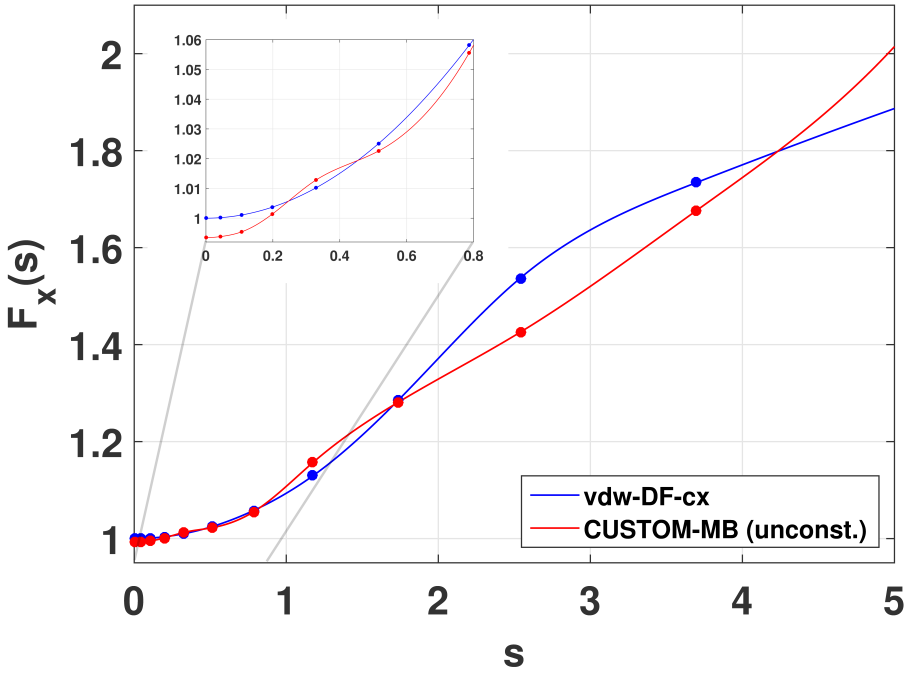


(a)

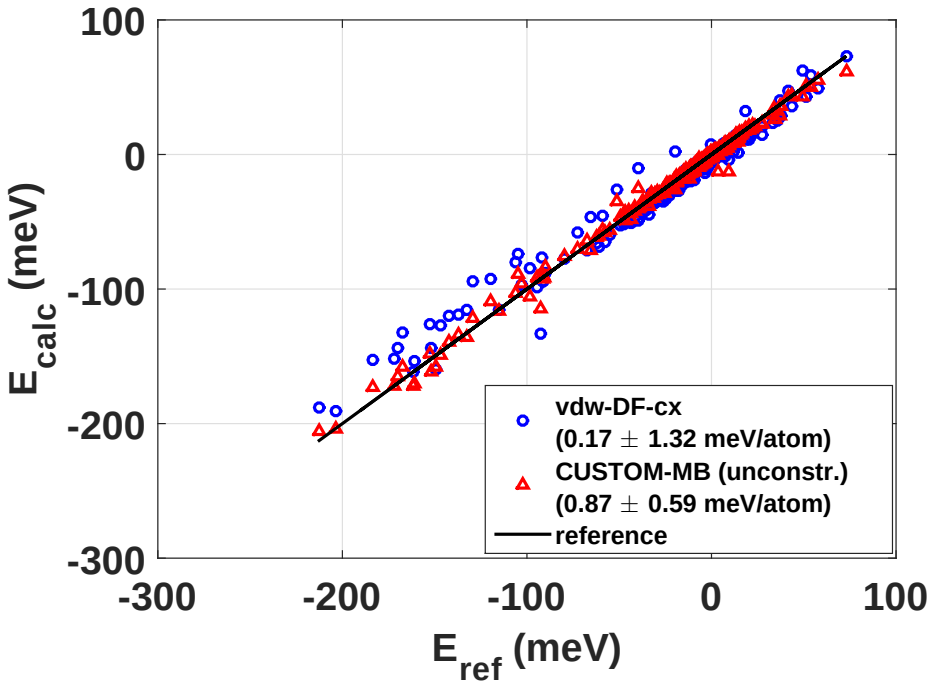


(b)

Fig. 4.18 The plots in (a) and (b) are the same as those in Figures 4.17a and 4.17b, with the exception that Bayesian constraints were used to fit the functional CUSTOM-3B.

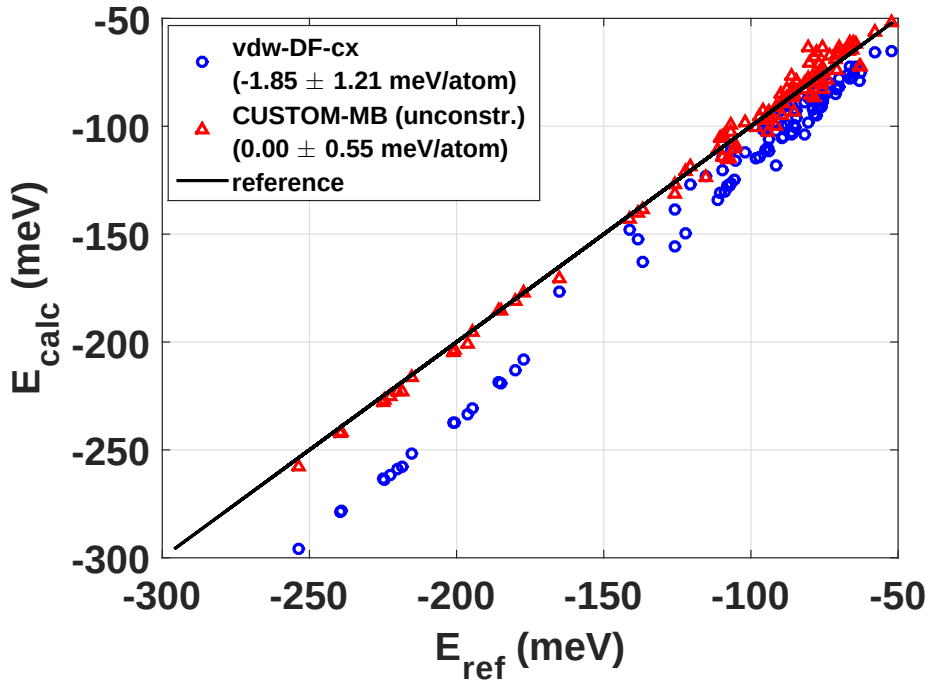

F_x(s)
s
vdw-DF-cx
CUSTOM-MB (unconst.)

(a)


E_calc (meV)
E_ref (meV)
vdw-DF-cx
(0.17 ± 1.32 meV/atom)
CUSTOM-MB (unconstr.)
(0.87 ± 0.59 meV/atom)
reference

(b)

Fig. 4.19



(c)

Fig. 4.19 (a) GGA exchange enhancement factor $F_x(s)$, optimized without constraints to reproduce 2-body and 3-body energies calculated with the MB-pol potential [6] of 330 water dimers and 100 water trimers, respectively (CUSTOM-MB (unconstr.), red) and that of vdW-DF-cx (LV-PW86R, blue), (b) 2-body energies of 330 dimers and (c) 3-body energies of 100 trimers calculated with the CUSTOM-1B functional (red) and with the vdW-DF1-cx functional (blue) against the MB-pol reference energies. The average errors ± RMS errors are shown in the legends.

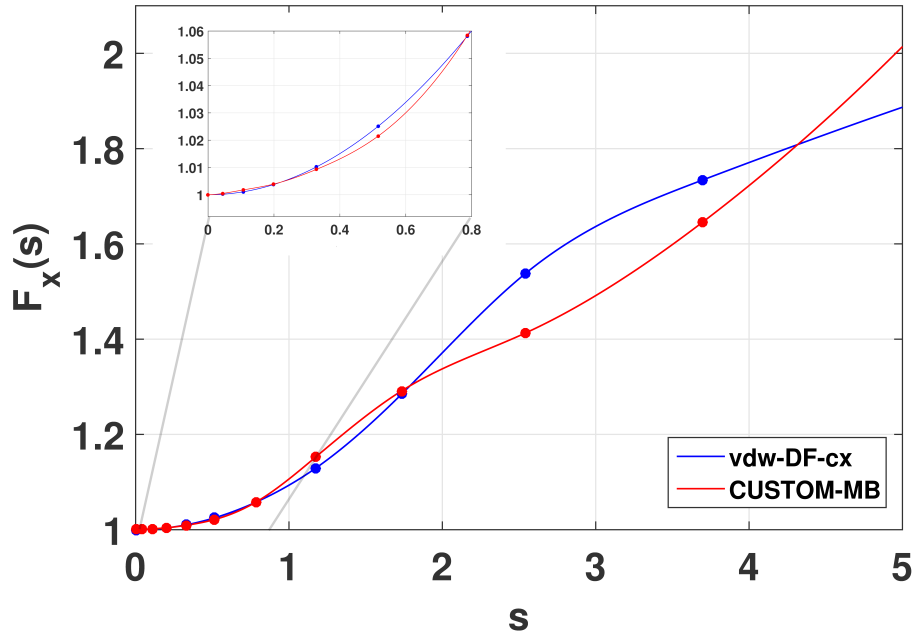
energies, and $\Delta E_n = 0.5$ meV for the set of 3-body energies. The optimized $F_x(s)$, labelled CUSTOM-MB (unconstr.), where MB represents "Many-body", is plotted in Figure. 4.19a, along with the LV-PW86R enhancement factor. In this case, without Bayesian constraints, the optimized $F_x(s)$ does not have any alarming deviations from the basic constraints, but small undulations appear for small values of s (see inset of Fig. 4.19a). With CUSTOM-MB (unconstr.), the average error of the 2-body energies is reduced to -0.09 $\pm$ 0.59 meV/atom (Fig. 4.19b), and the average error of the 3-body energies is reduced to 0.00 $\pm$ 0.55 meV/atom (Fig. 4.19c).

The plot in Fig. 4.20a shows $F_x(s)$ optimized for a dataset of 2-body energies of 300 dimers and the 3-body energies of 100 trimers, with Bayesian constraints (CUSTOM-MB), along with the LV-PW86R enhancement factor. Here Eq. 4.10 was maximized with $\Delta E_n = 0.3$ meV for the set of 2-body energies and $\Delta E_n = 0.5$ meV for the set of 3-body energies. Again, the optimization with constraints yields a functional that is monotonically increasing and that satisfies the HEG limit. With CUSTOM-MB, the average error of the dimer 2-body energies with respect to MB-pol is reduced to 0.02 $\pm$ 0.62 meV/atom (Fig. 4.20b) and the RMS error of the trimer 3-body energies is reduced to 0.00 $\pm$ 0.60 meV/atom (Fig. 4.20c). The dimer 2-body energies are worse compared to those calculated with CUSTOM-2B, but this comes with the advantage of improving the trimer 3-body energies.

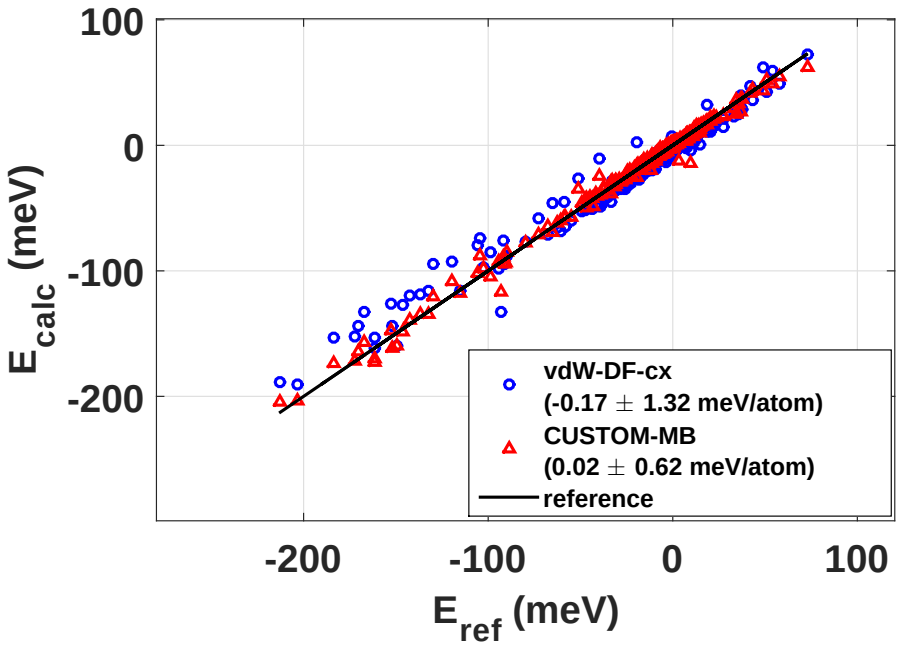
Although we have not included the 1-body energies in the optimization of the CUSTOM-MB functional, we studied the energies of the monomer calculated with this functional and compared them with the energies calculated with vdW-DF1-cx (Fig. 4.20d). We found that the RMS error of the 1-body energies of the CUSTOM-MB functional are relatively unimproved.

4.2.5 1-body error corrections

The 1-body, 2-body, and 3-body mean energy errors per atom and RMS fluctuation of the error with each CUSTOM functional previously described are summarized in Table 4.1. Overall, the optimization of the 1-body errors alone causes the 2-body and 3-body energies to become worse, as expected based on the analysis of density changes between the DFT and MB-pol equilibrium monomer geometry. Optimization of the 2-body energies or 3-body energies alone does not cause a significant change in the 1-body energies, but it does result in a poorer evaluation of the 3-body energies or 2-body energies, respectively. Optimization of both the

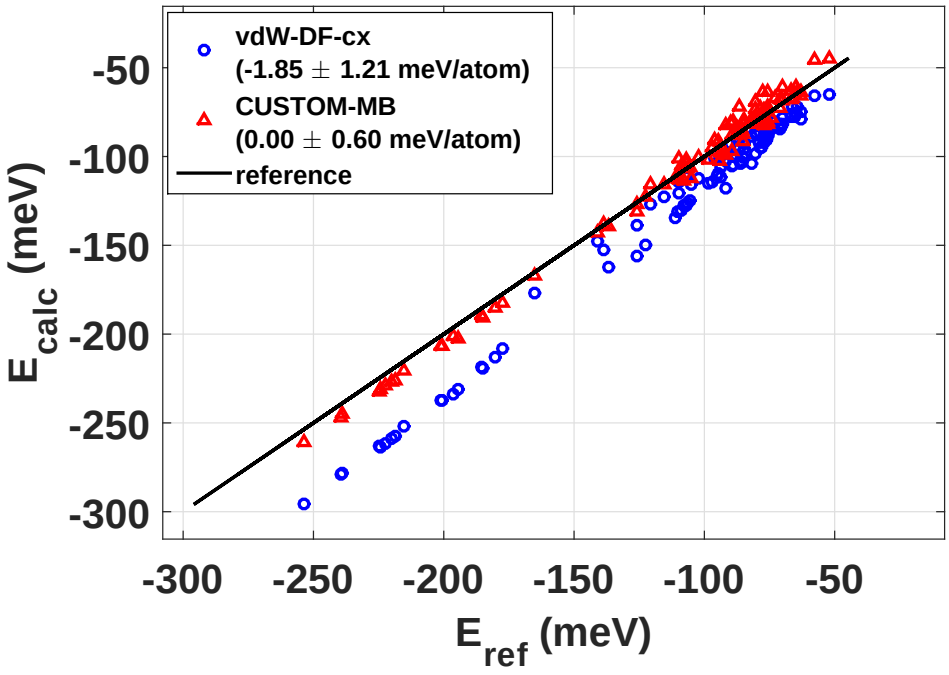

F_x(s)
s
vdw-DF-cx
CUSTOM-MB

(a)

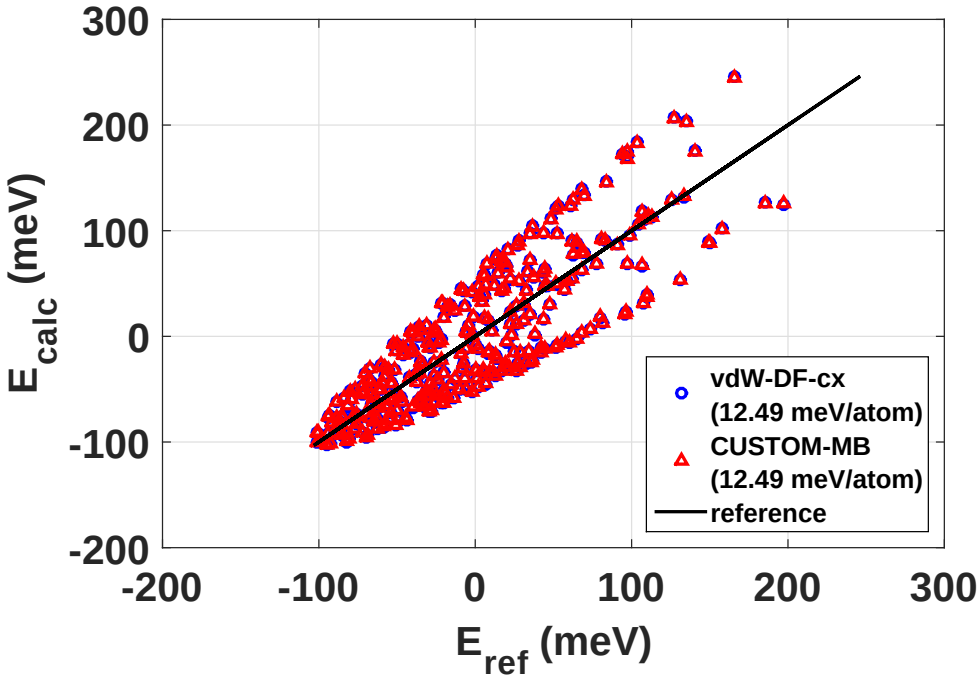

E_calc (meV)
E_ref (meV)
vdW-DF-cx
(-0.17 ± 1.32 meV/atom)
CUSTOM-MB
(0.02 ± 0.62 meV/atom)
reference

(b)

Fig. 4.20



(c)



(d)

Fig. 4.20 The plots in (a), (b), and (c) are the same as those in Figures 4.19a,4.19c and 4.17b, respectively with the exception that Bayesian constraints were used to fit the functional CUSTOM-MB and (d) 100 monomer energies calculated with the CUSTOM-MB functional (red) and with the vdW-DF1-cx functional (blue) plotted against the MB-pol reference energies. The average errors ± RMS errors are shown in the legends of (a)-(c), and the RMS errors are shown in the legend of (d).

2-body and 3-body energies causes an improvement in both of these terms without any significant changes to the 1-body energies.

Because of the difficulty involved in the simultaneous optimization of the 1-body energies with 2- and 3-body energies, we decided to focus on the optimization of only the 2- and 3-body energies with constrained DPPS, for application to calculations of water system properties. However, we also wanted to take into account the distortion energies of the water monomers to get the best possible description of water. To do this, we used the approach of Taylor et al. [31] to correct the 1-body energy terms after the optimization of 2- and 3-body terms, using the following equation:

$$E(DFT-\Delta_1) = E(DFT) + E^1(ref) - E^1(DFT), \tag{4.25}$$

Here $E(DFT-\Delta_1)$ is the total energy of a system of water molecules, including 1-body energy corrections. $E^1(ref)$ is the sum of all the reference 1-body terms of the system, calculated by MB-pol. To correct the total DFT energy $E(DFT)$, $E^1(ref)$ is added to it, and the total DFT-calculated 1-body energy $E^1(DFT)$ is subtracted. In the following, the functionals that have been corrected this way are referred to as their original name with -Δ_1 added to the end. For example, the 1-body-corrected version of CUSTOM-MB becomes CUSTOM-MB-Δ_1.

As a test, we analyzed the predictions of the equilibrium geometries and electric dipole moments of water monomers with the Bayesian constrained CUSTOM functionals, including a 1-body corrected version of CUSTOM-2B and CUSTOM-MB. The results of these functionals are compared with the results of vdW-DF-cx and MB-pol in Table 4.2.

The CUSTOM-2B and CUSTOM-MB functionals that do not include 1-body energy corrections show small improvements overall from vdW-DF-cx in predicting monomer equilibrium geometries. The CUSTOM-1B functional gives the best geometry of the unconstrained functionals, improving the O-H distances by 1.1 % and keeping the same error in the H-O-H angle. The CUSTOM-1B functional, however, gives the largest error in the electric dipole moment, increasing it by 1.8%. The electric dipole moments predicted by the other CUSTOM functionals are slightly better, with the error reduced by 0.4% for CUSTOM-2B and increased by 0.2% with CUSTOM-MB with respect to the dipole moment error predicted by vdW-DF-cx. When taking 1-body energy corrections into account, both the CUSTOM-2B-Δ_1

Table 4.1 Mean errors of 300 1-body energies, 330 2-body energies, and 100 3-body energies calculated with the vdW-DF-cx functional and all CUSTOM functionals described here. The monomers, dimers, and trimers are those of the sets used in the optimizations. Each entry gives the mean deviation of the energy from the reference energy $\pm$ the RMS fluctuation of the energy deviation per atom. In the case of the 1-body terms, only the RMS values are included, since the average energy error for each functional is $\equiv$ 0. All values are in meV/atom.

Functional	1-body	2-body	3-body
vdW-DF-cx	$\pm$ 12.49	-0.17 $\pm$ 1.32	-1.85 $\pm$ 1.21
CUSTOM-1B (unconstr.)	$\pm$ 0.15	-0.65 $\pm$ 1.11	1.23 $\pm$ 23.80
CUSTOM-1B	$\pm$ 0.59	9.30 $\pm$ 22.96	6.00 $\pm$ 21.47
CUSTOM-2B (unconstr.)	$\pm$ 5.91	-0.08 $\pm$ 0.36	3.61 $\pm$ 45.93
CUSTOM-2B	$\pm$ 9.03	-0.01 $\pm$ 0.48	3.35 $\pm$ 6.02
CUSTOM-3B (unconstr.)	$\pm$ 12.60	-1.49 $\pm$ 2.35	0.00 $\pm$ 0.27
CUSTOM-3B	$\pm$ 12.55	-0.86 $\pm$ 2.11	0.00 $\pm$ 0.31
CUSTOM-MB (unconstr.)	$\pm$ 12.57	-0.09$\pm$ 0.60	0.00 $\pm$ 0.55
CUSTOM-MB	$\pm$ 12.49	0.02 $\pm$ 0.62	0.00 $\pm$ 0.60

and CUSTOM-3B-Δ_1 improve the monomer geometry with respect to their original versions, bringing them into almost perfect agreement with MB-pol.

4.3 Conclusions

In this chapter, the basic concepts involved in the DPPS method were explained in terms of the fitting of an inter-atomic potential and the fitting of a GGA XC functional to a set of reference energies. A modified version of DPPS, which uses Bayesian probability theory to include constraints on the parameters, was also presented.

Although the DPPS method is a general method that can in principle be applied to many types of data and functionals, so far we have only applied it to the optimization of energies by parameterizing GGA exchange functionals. We have determined with a simple test that DPPS is able to correctly optimize GGA exchange by starting with the PBE form of $F_x(s)$ and reproducing the form of PW86R. After confirming that DPPS is effective in optimizing GGA exchange, we went through the process of optimization of a GGA exchange functional to datasets of water 1-, 2-, and 3-body energies predicted by the MB-pol potential. The optimizations to 1-, 2-, and 3-body

Table 4.2 Equilibrium distances between O and H r_{OH}, equilibrium angle between two O-H bonds, and electric dipole moment at the reference equilibrium geometry calculated with MB-pol of a water monomer, calculated with the vdW-DF-cx functional and CUSTOM functionals optimized with Bayesian constrained DPPS. To the right of each calculated values, the percent errors of values taken with respect to the values of MB-pol are displayed. In all calculations, a 300 Ry mesh cutoff, 20 Å lattice constant, and $^{(s)}q\zeta$-dp basis set were used.

Method	r_{OH} (Å)		$\theta_{HOH}(^\circ)$		Electric dipole (D)	
	calc.	% error	calc.	% error	calc.	% error
MB-pol	0.9578	-	104.51	-	1.864	-
vdW-DF-cx	0.9707	1.3	104.35	-0.2	1.813	-2.8
CUSTOM-1B	0.9600	0.2	104.69	0.2	1.780	-4.6
CUSTOM-2B	0.9669	1.0	104.76	0.2	1.820	-2.4
CUSTOM-2B-Δ_1	0.9580	0.0	104.51	0.0	1.820	-2.4
CUSTOM-MB	0.9706	1.3	104.17	-0.3	1.809	-3.0
CUSTOM-MB-Δ_1	0.9582	0.0	104.52	0.0	1.809	-3.0

energy datasets were done for each set separately, both with unconstrained and constrained DPPS. Both the constrained and unconstrained methods improve the dataset energies, with the unconstrained fit showing the most improvement due to the extra flexibility in functional form. The unconstrained method shows exactly how the functional needs change to reproduce the reference energies while disregarding physical constraints, and the constrained method shows the best fit that we can achieve while obeying these constraints and predicts more practical functionals.

An analysis of the optimization of GGA exchange for 1-body energies and of electron densities in this chapter revealed the physical root of the inability of GGA exchange functionals to correct self-interaction errors, specifically in the case of water. We found that the density changes corresponding to the shifting of atoms in a water monomer to match a reference geometry have little dependence on s, making it impossible for a function $F_x(s)$ that only depends on s to reflect the changes in energy needed to correct the 1-body energies.

Because of the poor ability of the GGA exchange to reproduce the 1-body energies, we decided to focus on the simultaneous optimization of 2- and 3-body energies, and add a 1-body energy correction term to the total energy to include the effects of monomer deformation energies in our studies involving larger water sys-

tems. Considering the errors in 1-, 2-, and 3-body terms, the functional optimized with constraints to both 2- and 3-body energies that includes 1-body corrections (CUSTOM-MB-Δ_1) performs best overall of all of the optimized functionals. The next step in the analysis of these functionals is to test them on larger water systems, ranging from small clusters to bulk water, which will be discussed in the following chapters.

Chapter 5

The Thermodynamic Integration (TI) method for AIMD convergence testing

Before using the new DPPS-optimized functionals that were presented in Chapter 4 for *ab intio* Molecular Dynamics (AIMD) simulations of liquid water, it was important to perform tests to determine the simulation parameters for which the calculated structural and thermodynamic properties are adequately converged. However, there are a few difficulties in convergence testing for AIMD simulations. One problem is that the simulations are subject to large statistical errors in the structural and thermodynamic properties. Because of this, the convergence of these parameters can be masked. Another issue is that several different simulations are required in order to test all of the parameters, and performing sufficiently long simulations in order to get enough statistical data is very computationally expensive. In this chapter, a method that we developed as a way to avoid these problems is presented. This method, which we refer to as the Thermodynamic Integration (TI) method, evaluates parameter convergence without multiple, costly AIMD simulations. Instead, the TI method requires the evaluation of only a small set of trajectories from an AIMD simulation. Then, in the spirit of thermodynamic integration approaches, the average change of a given property is estimated by repeating calculations with the same geometries and different parameter values.

In the following, a study of the convergence of liquid water properties using both methods is presented. First, results from the "old" method, which requires

repetitions of full AIMD simulations, is presented. These results are then compared with those of the TI method. The convergence of the following parameters was tested:

- *Basis set*. Different types of basis sets, and how they affect the convergence of results is discussed in Section 2.6.2.
- *Mesh cutoff*. This determines the fineness of the real-space grid used for integration.
- *Density Matrix (DM) tolerance*. This determines the cutoff of the maximum difference between the output and the input on each element of the DM in a SCF cycle.
- *Number of molecules in one simulation cell*.

Changing these parameters, the convergence of the following properties was studied:

- *Average simulation pressure* $\langle P \rangle$
- *Average order parameter* $\langle \zeta \rangle$. The order parameter ζ was introduced by Russo and Tanaka [112], and is intended to measure the local translational order in the second shell of neighbours in liquid water. For a water molecule, ζ is defined as the difference between the distance to its first neighbor not considered hydrogen bonded $r^{1^{st}}_{non-H-bonded}$ and the distance to the last neighbor that is hydrogen bonded $r^{last}_{H-bonded}$. At lower temperatures, when water is closer to a crystalline state, the average values of ζ have higher values (higher translational order), and at lower temperatures they transition to lower values (lower translational order) [112].
- *The heights of the first maximum* $g_{OO}^{1^{st}max}$ *and the first minimum* $g_{OO}^{1^{st}min}$ *in the oxygen-oxygen radial distribution function (RDF)* $g_{OO}(r)$*, where* r *is the distance between the oxygen atoms .*

5.1 Convergence testing with AIMD simulations

In this section, the results of what we refer to as the “old“ method of convergence testing are presented. Separate AIMD simulations are done for each change

in each of the parameters tested. For all AIMD simulation results presented in this chapter, equilibration was accomplished with temperature annealing (velocity rescaling), while the production runs were performed with Verlet integration at constant energy [3], and each time step was 0.5 fs. All of the simulations were equilibrated at a temperature of 300 K during 5 ps, and after equilibration each one was continued for 10 ps of a production run. Both annealing and production runs were performed at a constant volume (fixed cell size and shape, under periodic boundary conditions).

5.1.1 Property dependence on average simulation temperature

Due to the fluctuations in temperature in the simulations, the initial configuration of the production run (which is the final configuration of the annealing part) can have a temperature that is higher or lower that the target average temperature. This is a typical problem that can cause simulations with the same target temperature to have differing values of average temperature $\langle T \rangle$ (defined such that $\langle E_{kin} \rangle = 3k\langle T \rangle$, where $\langle E_{kin} \rangle$ is the average kinetic energy and k is the Boltzmann constant) over the production runs. Unfortunately, these differences in temperature affect the properties that we are testing for convergence. Ideally, all of the simulations should have exactly the same value of $\langle T \rangle$.

In order to estimate the dependence of the values of $\langle P \rangle$, $\langle \zeta \rangle$, $g_{OO}^{1^{st}max}$ and $g_{OO}^{1^{st}min}$ on temperature, AIMD simulations with different target temperatures were performed, using a $^{(p)}$dζ+p basis set, 32 water molecules, and a 200 Ry mesh cutoff (Table 5.3, Fig. 5.1). With an estimation of how each property changes with temperature, the resulting magnitudes for convergence testing can be adjusted accordingly, using to the following formula:

$$\alpha_f = \alpha_i + (T_f - T_i)\frac{d\alpha}{dT}, \tag{5.1}$$

where α_i is the average value $\langle \alpha \rangle$ of the calculated property $\alpha = \left\{ \langle P \rangle, \langle \zeta \rangle, g_{OO}^{1^{st}max}, g_{OO}^{1^{st}min} \right\}$, α_f is its temperature-adjusted value, T_i is the average temperature $\langle T \rangle$, T_f is the desired target temperature, and $\frac{d\alpha}{dT}$ is the slope of the fit of α against $\langle T \rangle$, as shown in Fig. 5.1 and Table 5.3.

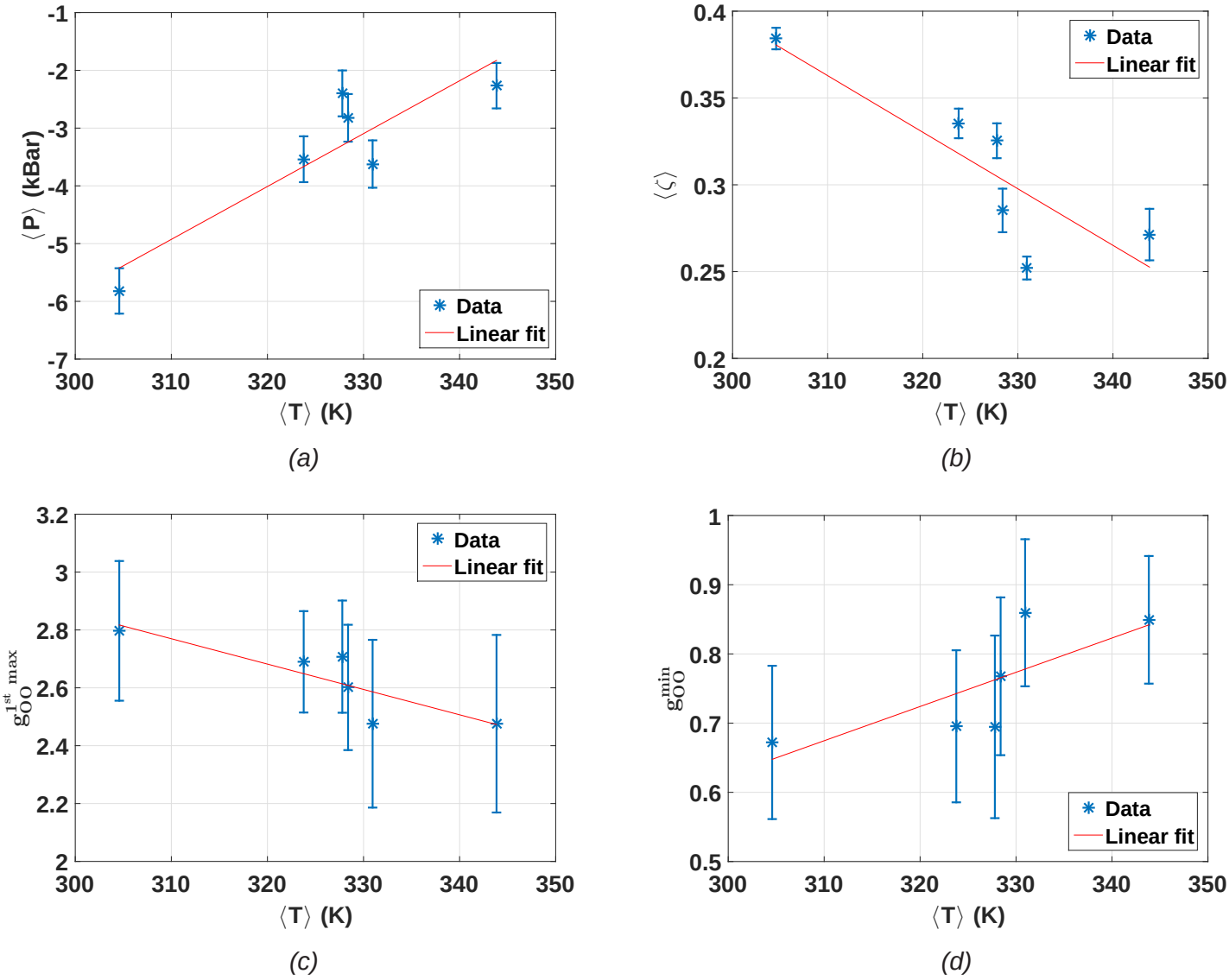


Fig. 5.1 Dependence of (a) the average simulation pressure $\langle P \rangle$, (b) the average order parameter $\langle \zeta \rangle$, and the position of (c) the first maximum $g_{OO}^{1^{st}max}$ and (d) the first minimum $g_{OO}^{1^{st}min}$ in the oxygen-oxygen RDF on average simulation temperature $\langle T \rangle$. All simulations were done with the vdW-DF-cx functional, 32 water molecules per simulation cell, the ${}^{(p)}d\zeta$ + p the basis set, a 200 Ry mesh cutoff, and a 5×10^{-4} Bohr^{-3} DM tolerance. The error bars were calculated using the first term of Eq. 5.2, and the slopes of the linear fits are given in Table 5.1.

Table 5.1 Values from AIMD liquid water simulations of the average simulation pressure $\langle P \rangle$, the average order parameter $\langle \zeta \rangle$, and the position of the first maximum $g_{OO}^{1^{st}max}$ and the first minimum $g_{OO}^{1^{st}min}$ in the oxygen-oxygen RDF $g_{OO}(r)$. All simulations were done with the vdW-DF-cx functional, 32 water molecules per unit cell, the $^{(p)}d\zeta$ + p basis set, a 300 Ry mesh cutoff, and a 5×10^{-4} Bohr^{-3} DM tolerance. These numbers are used to calculate the linear fit of each property plotted against the average simulation temperature $\langle T \rangle$ (Fig. 5.1). The slopes $d\alpha/d\langle T \rangle$ of the fits are included in the bottom table.

$\langle P \rangle (kBar)$	$\langle T \rangle (K)$	$\langle \zeta \rangle$	$g_{OO}^{1^{st}max}$	$g_{OO}^{1^{st}min}$
-5.8 $\pm$ 0.4	305 $\pm$ 1	0.38 $\pm$ 0.01	2.80 $\pm$ 0.12	0.67 $\pm$ 0.07
-3.5 $\pm$ 0.4	324 $\pm$ 1	0.34 $\pm$ 0.01	2.70 $\pm$ 0.12	0.70 $\pm$ 0.11
-2.4 $\pm$ 0.4	328 $\pm$ 2	0.33 $\pm$ 0.01	2.71 $\pm$ 0.22	0.70 $\pm$ 0.13
-2.8 $\pm$ 0.4	328 $\pm$ 6	0.29 $\pm$ 0.01	2.60 $\pm$ 0.09	0.77 $\pm$ 0.05
-3.6 $\pm$ 0.4	330 $\pm$ 4	0.25 $\pm$ 0.01	2.48 $\pm$ 0.15	0.86 $\pm$ 0.11
-2.3 $\pm$ 0.4	344 $\pm$ 5	0.27 $\pm$ 0.01	2.48 $\pm$ 0.17	0.85 $\pm$ 0.06

$\frac{d\langle P \rangle}{d\langle T \rangle}$	$\frac{d\langle \zeta \rangle}{d\langle T \rangle}$	$\frac{dg_{OO}^{1^{st}max}}{d\langle T \rangle}$	$\frac{dg_{OO}^{1^{st}min}}{d\langle T \rangle}$
0.092	-0.0032	-0.0088	0.0049

5.1.2 AIMD results

Using Eq. 5.1, temperature adjusted values of the average simulation pressure $\langle P \rangle$, order parameter $\langle \zeta \rangle$, and positions of the first maximum $g_{OO}^{1^{st}max}$ and the first minimum $g_{OO}^{1^{st}min}$ of the oxygen-oxygen RDF are estimated for the same value of T and compared. The results for each convergence tests are summarized in Table 5.2 and Figs. 5.5-5.8 of Section 5.2.

In the convergence test for the number of molecules per unit cell, basis set, and mesh cutoff, values were adjusted for T=311 K as needed, since this was the average simulation temperature for the reference values (first row of each section in Table 5.2). In the DM tolerance convergence test, for which a mesh cutoff of 300 Ry rather than 200 Ry was used, values were adjusted for T=323 K as needed.

Table 5.2 Average temperature $\langle T \rangle$ of AIMD simulations performed for convergence tests, along with temperature adjusted (Eq. 5.1) values of the following properties: average simulation pressure $\langle P \rangle$, average order parameter $\langle \zeta \rangle$ and the positions of the first maximum and minimum in the oxygen-oxygen RDF $g_{OO}(r)$. All simulations were done with the vdW-DF-cx functional, and unless otherwise noted, 32 molecules per unit cell, the $^{(p)}$dζ+p basis set, a 200 Ry mesh cutoff, and a 5×10^{-4} Bohr^{-3} DM tolerance. All values in the tests for number of molecules, basis set, and mesh cutoff are adjusted for T=311 K, and in the tests for DM tolerance they are adjusted for T=323 K. All temperatures are in K, pressures are in kBar, and DM tolerances are in Bohr^{-3}.

Number of molecules ($T = 311K$)

N	$\langle T \rangle$	$\langle P \rangle$	$\langle \zeta \rangle$	$g_{OO}^{1^{st}max}$	$g_{OO}^{1^{st}min}$
32	311 $\pm$ 5	-3.4 $\pm$ 0.9	0.38 $\pm$ 0.02	3.04 $\pm$ 0.06	0.45 $\pm$ 0.05
64	328 $\pm$ 4	-4.5 $\pm$ 0.7	0.39 $\pm$ 0.02	2.86 $\pm$ 0.06	0.65 $\pm$ 0.05
128	323 $\pm$ 3	-5.1 $\pm$ 0.5	0.43 $\pm$ 0.01	2.83 $\pm$ 0.06	0.62 $\pm$ 0.05

Basis set ($T = 311K$)

Basis set	$\langle T \rangle$	$\langle P \rangle$	$\langle \zeta \rangle$	$g_{OO}^{1^{st}max}$	$g_{OO}^{1^{st}min}$
$^{(p)}$dζ+p	311 $\pm$ 5	-3.4 $\pm$ 0.9	0.38 $\pm$ 0.02	3.04 $\pm$ 0.06	0.45 $\pm$ 0.05
tζ+dp	302 $\pm$ 1	0.2 $\pm$ 0.5	0.38 $\pm$ 0.01	2.87 $\pm$ 0.06	0.56 $\pm$0.05
$^{(s)}$qζ+dp	318 $\pm$ 2	-0.8 $\pm$ 0.5	0.49 $\pm$ 0.01	3.03 $\pm$ 0.06	0.49 $\pm$0.05
qζ+dp	313 $\pm$ 5	1.8 $\pm$ 1.0	0.48 $\pm$ 0.03	3.07 $\pm$ 0.06	0.51 $\pm$0.05

Mesh cutoff ($T = 311K$)

Mesh cutoff (Ry)	$\langle T \rangle$	$\langle P \rangle$	$\langle \zeta \rangle$	$g_{OO}^{1^{st}max}$	$g_{OO}^{1^{st}min}$
200	311 $\pm$ 5	-3.4 $\pm$ 0.9	0.38 $\pm$ 0.02	3.04 $\pm$ 0.06	0.45 $\pm$ 0.05
300	314 $\pm$ 5	-4.0 $\pm$ 1.0	0.38 $\pm$ 0.03	2.78 $\pm$ 0.06	0.62 $\pm$ 0.05
400	315 $\pm$ 2	-2.7 $\pm$ 2.0	0.49 $\pm$ 0.01	2.71 $\pm$ 0.06	0.64 $\pm$ 0.05

DM tolerance (300 Ry Mesh cutoff, $T = 323K$)

DM tol.	$\langle T \rangle$	$\langle P \rangle$	$\langle \zeta \rangle$	$g_{OO}^{1^{st}max}$	$g_{OO}^{1^{st}min}$
2$\times 10^{-4}$	323 $\pm$ 5	-3.5 $\pm$ 0.4	0.35 $\pm$ 0.01	2.81 $\pm$ 0.06	0.63 $\pm$ 0.05
1$\times 10^{-4}$	314 $\pm$ 5	-3.5 $\pm$ 0.4	0.42 $\pm$ 0.03	2.81 $\pm$ 0.06	0.63 $\pm$ 0.05
5$\times 10^{-5}$	324 $\pm$ 2	-3.5 $\pm$ 0.4	0.34 $\pm$ 0.01	2.69 $\pm$ 0.06	0.70 $\pm$ 0.05
2$\times 10^{-5}$	323 $\pm$ 1	-3.7 $\pm$ 0.6	0.31 $\pm$ 0.01	2.60 $\pm$ 0.06	0.73 $\pm$ 0.05
1x10^{-5}	327 $\pm$ 2	-3.5 $\pm$ 0.6	0.34 $\pm$ 0.01	2.78 $\pm$ 0.06	0.65 $\pm$ 0.05

Error bars of each property α that was temperature-adjusted were calculated as

$$\delta_\alpha = \frac{\sigma_\alpha}{\sqrt{\frac{\tau_{max}}{\tau_\alpha}}} + \frac{d\alpha}{dT}\frac{\sigma_T}{\sqrt{\frac{\tau_{max}}{\tau_T}}}. \tag{5.2}$$

Here τ_{max} is the simulation time over which the property was averaged, τ_α is the autocorrelation time of the property α, and τ_T is the autocorrelation time of the temperature. The values σ_T and σ_α are the standard deviations of the property α and simulation temperature, respectively. For example, σ_α is calculated as follows:

$$\sigma_\alpha = \sqrt{\frac{1}{\tau_{max}}\int_0^{\tau_{max}} (\alpha(\tau) - \langle\alpha\rangle)^2 d\tau}, \tag{5.3}$$

with

$$\langle\alpha\rangle = \frac{1}{\tau_{max}}\int_0^{\tau_{max}} \alpha(\tau) d\tau. \tag{5.4}$$

For properties that did not need to be temperature-adjusted, the error bars were calculated using only the first term of Eq. 5.2.

The autocorrelation function of each property was estimated as

$$C_\alpha(t) = \frac{1}{(\tau_{max} - t)\sigma_\alpha^2}\int_0^{\tau_{max}-t} (\alpha(\tau) - \langle\alpha\rangle)(\alpha(\tau + t) - \langle\alpha\rangle) d\tau. \tag{5.5}$$

The autocorrelation time τ_α was then estimated by fitting $C_\alpha(t)$ to an exponential function such that $C_\alpha(t) \propto e^{-t/\tau_\alpha}$. This same method was used to find the temperature autocorrelation time τ_T for each simulation.

5.2 The Thermodynamic Integration method

5.2.1 Methodology

In this section, a new method of convergence testing for AIMD parameters, referred to as the Thermodynamic Integration (TI) method, is presented. With this method, we can significantly reduce computation time by using only a small percentage of snapshots from an AIMD trajectory to estimate simulation properties, rather than running an entire simulation.

In the TI method, we consider the linear mix of two different potentials, V_0 and V_1, which can be used to define a new hybrid potential

$$V_\lambda = (1-\lambda)V_0 + \lambda V_1. \tag{5.6}$$

Here $0 < \lambda < 1$, V_0 is the potential used in the AIMD simulation, and V_1 is a new potential for which we want to estimate certain properties. V_1 can be a potential with an entirely different functional from V_0, or one with the same functional form, but different parameters, such a different basis set, mesh cutoff, etc. The derivative of the potential is

$$\frac{dV_\lambda}{d\lambda} = V_1 - V_0 = \Delta V. \tag{5.7}$$

Considering a number of snapshots described by the hybrid potential V_λ, we can evaluate the average of the magnitude α as a function of λ. The ensemble average of α is

$$\langle\alpha\rangle_\lambda = \frac{1}{Z_\lambda}\int d^{3N}x\, e^{-\beta V_\lambda(x)}\alpha_\lambda(x), \tag{5.8}$$

where

$$Z_\lambda = \int d^{3N}x\, e^{-\beta V_\lambda(x)} \tag{5.9}$$

is the partition function. Here $\beta = \frac{1}{kT}$, with k equal to the Boltzmann constant, T is the temperature, and $\alpha_\lambda(x)$ is the property calculated with the potential $V_\lambda(x)$. The term x represents the coordinates of all of the N atoms of the set of snapshots. The derivative of the ensemble average is

$$\begin{aligned}\frac{d\langle\alpha\rangle_\lambda}{d\lambda} = \frac{1}{Z_\lambda}\int d^{3N}x\frac{\partial\alpha_\lambda(x)}{\partial\lambda}e^{\beta V_\lambda(x)} - \frac{\beta}{Z_\lambda}\int d^{3N}x\, e^{-\beta V_\lambda(x)}\alpha_\lambda(x)\Delta V(x) \\ + \frac{\beta}{(Z_\lambda)^2}\int d^{3N}x\, e^{-\beta V_\lambda(x)}\Delta V(x)\int d^{3N}x\, e^{-\beta V_\lambda(x)}\alpha_\lambda(x) = \\ = \langle\frac{\partial\alpha}{\partial\lambda}\rangle_\lambda - \beta[\langle\alpha\Delta V\rangle_\lambda - \langle\alpha\rangle_\lambda\langle\Delta V\rangle_\lambda].\end{aligned} \tag{5.10}$$

The change in the a property α can then approximated by a linear extrapolation from $\lambda = 0$:

$$\begin{aligned}\langle\Delta\alpha\rangle = \langle\alpha\rangle_{\lambda=1} - \langle\alpha\rangle_{\lambda=0} \simeq (\frac{d\langle\alpha\rangle_\lambda}{d\lambda})_{\lambda=0} = \\ \langle\frac{\partial\alpha}{\partial\lambda}\rangle_{\lambda=0} - \beta\left[\langle\alpha\Delta V\rangle_{\lambda=0} - \langle\alpha\rangle_{\lambda=0}\langle\Delta V\rangle_{\lambda=0}\right].\end{aligned} \tag{5.11}$$

Here the averages $\langle\ \rangle_{\lambda=0}$ are calculated for a subset of the configurations (snapshots) from one AIMD simulation with the potential V_0, i.e., with $\lambda = 0$. The interval between these configurations is on the order of the autocorrelation time for the studied property. Since this correlation time is typically two orders of magnitude larger than the MD time step, estimating $\langle\Delta\alpha\rangle$ with Eq. 5.11 is much faster than repeating the entire AIMD simulation with potential V_2. Furthermore, as V_1 and V_2 become very close, which always occurs near convergence, $\langle\alpha\rangle_{\lambda=1} - \langle\alpha\rangle_{\lambda=1}$ will become very small and easily masked by statistical noise while $\langle\Delta\alpha\rangle$ given by Eq. 5.11 will become increasingly accurate.

5.2.2 Testing of the TI method

Pressure (AIMD)

To test the accuracy of pressure estimates done with the TI method, we used it to estimate pressure changes in AIMD simulations of liquid water with different DFT functionals. We started with the vdW-DF-cx functional and estimated the changes in pressure for the following functionals: PBE, vdW-DF^{PBE}, which uses the same correlation functional as vdW-DF and PBE exchange, vdW-DF, and the Vydrov and Van Voorhis exchange-correlation functional (VV) [127].

Before the tests was performed, the change in $\Delta\langle P\rangle$ was estimated for different numbers of snapshots out of a 20,000 step (10 ps) trajectory of an AIMD simulation with the vdW-DF-cx functional, using 32 molecules per unit cell, the $^{(p)}$dζ+p basis set, a 300 Ry mesh cutoff, a 5×10^{-4} Bohr^{-3} DM tolerance, and a target temperature of 300 K. After testing the convergence of $\Delta\langle P\rangle$ with the number of trajectory steps, we found that only 200 snapshots were necessary to achieve a converged estimation of $\Delta\langle P\rangle$. Then, using a set of 200 snaphots from the 20,000 step trajectory, changes in average pressure $\Delta\langle P\rangle$ with the different functionals were estimated with respect to the average pressure of the full simulation. According to Eq. 5.11, these changes in pressure $\Delta\langle P\rangle$ can be estimated as

$$\Delta\langle P\rangle \equiv \langle P\rangle_{\lambda+\Delta\lambda} - \langle P\rangle_{\lambda} \simeq \langle\Delta P\rangle_{\lambda} - \beta[\langle P\Delta V\rangle_{\lambda} - \langle P\rangle_{\lambda}\langle P\Delta V\rangle_{\lambda}], \tag{5.12}$$

where $\langle\Delta P\rangle \equiv \langle P_{\lambda+\Delta\lambda} - P_{\lambda}\rangle$, $\langle\Delta V\rangle \equiv \langle V_{\lambda+\Delta\lambda} - V_{\lambda}\rangle$, and $\langle\ \rangle_{\lambda}$ indicates the average over a subset of snapshots of an AIMD trajectory with potential V_{λ}.

Because we used snapshots produced by the vdW-DF-cx functional and estimate changes in pressure with respect to this functional, the snapshot energies calculated with this functional are represented by V_0 ($\lambda = 0$). Then, each of the other functionals are used to estimate changes in pressure for $\lambda = 1$, with V_1 representing the snapshot energies calculated with each of these functionals.

The TI estimated values of $\langle P \rangle$ for all functionals are compared with the values averaged over AIMD simulations in Fig. 5.2. The AIMD value of $\langle P \rangle$ for the vdW-DF-cx functional was calculated using the same 20,000 step trajectory from which the 200 snapshots used in the TI estimated were extracted. The AIMD values of $\langle P \rangle$ for PBE and vdW-DF^{PBE} were taken from Ref. [20], which used 64 molecules per unit cell, and values of $\langle P \rangle$ of vdW-DF and VV were taken from Ref. [129], which used 200 molecules per unit cell. Simulations presented in both references were done with the $^{(p)}$dζ+p basis set.

It can be seen that, taking error bars into account, TI accurately predicts the values of $\langle P \rangle$ for both the PBE and vdW-DF^{PBE} functionals. Although error bars for the AIMD results for vdW-DF and VV are not included, a comparison with the TI results for these functionals suggests that the estimated pressures are less accurate for these functionals. This could be a result of a larger perturbation in these functionals with respect to vdW-DF-cx. As discussed previously in the classical MD case, this can cause the trajectory of original functional (vdW-DF-cx in this case) to poorly represent the new functional and result in less accurate pressure estimations.

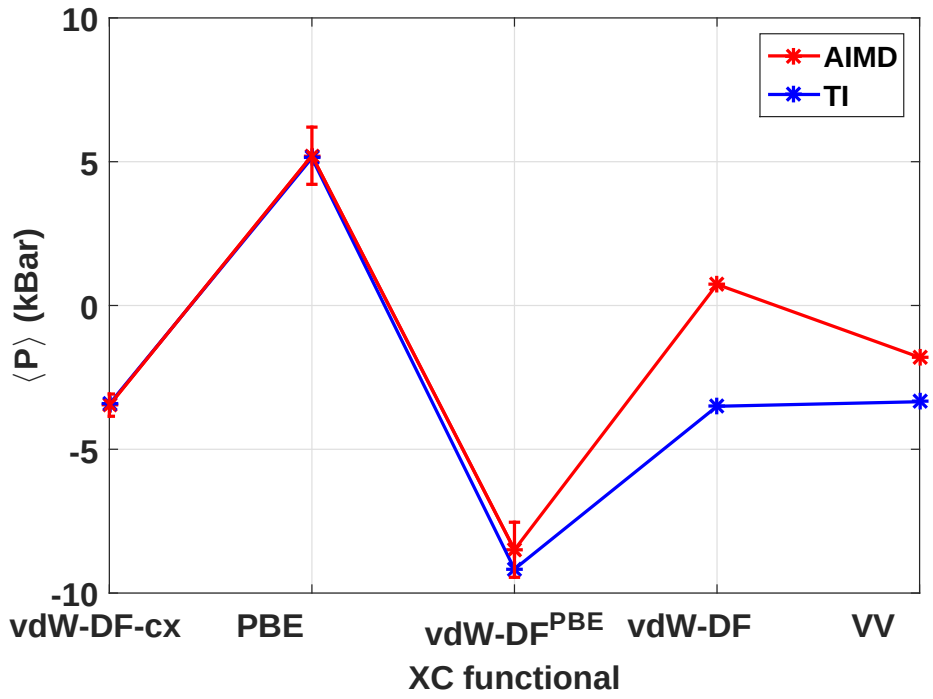


Fig. 5.2 Comparison of average pressures $\langle P \rangle$ over 20,000 step AIMD trajectories, with different functionals (red), and their TI estimates (blue, Eq. 5.12), using a single set of 200 snapshots produced by an AIMD simulation with the vdW-DF-cx functional, with 32 molecules per unit cell, the ${}^{(p)}$dp+p basis set, a 300 Ry mesh cutoff, and a 5×10^{-4} Bohr^{-3} DM tolerance. The AIMD value for vdW-DF-cx was averaged over the full trajectory of this simulation. The AIMD values and error bars for PBE and vdW-DF^{PBE} were taken from Ref. [20], which used 64 molecules per unit cell, and the values for vdW-DF and VV were taken from Ref. [129], which used 200 molecules per unit cell. In both references, the ${}^{(p)}$dζ+p basis set was used. Error bars were not provided in Ref. [129].

AIMD oxygen-oxygen radial distribution function: changes with basis set

The change between the radial distribution functions (RDFs) $\Delta\langle g_{\alpha\beta}(r)\rangle$ of species α and β, at a separation r, between two potentials V_0 and V_1, estimated with the TI method is

$$\Delta\langle g_{\alpha\beta}(r)\rangle = -\beta\left[\langle g_{\alpha\beta}(r)\Delta V\rangle_{\lambda=0} - \langle g_{\alpha\beta}(r)\rangle\langle \Delta V\rangle_{\lambda=0}\right]. \tag{5.13}$$

In this case, the first term of Eq. 5.11 is zero. The same set of snapshots is used for each value of λ, and the same geometries yields the same $g_{\alpha\beta}$. Thus, $\langle \frac{\partial g_{\alpha\beta}}{\partial \lambda} \rangle_{\lambda=0} = 0$.

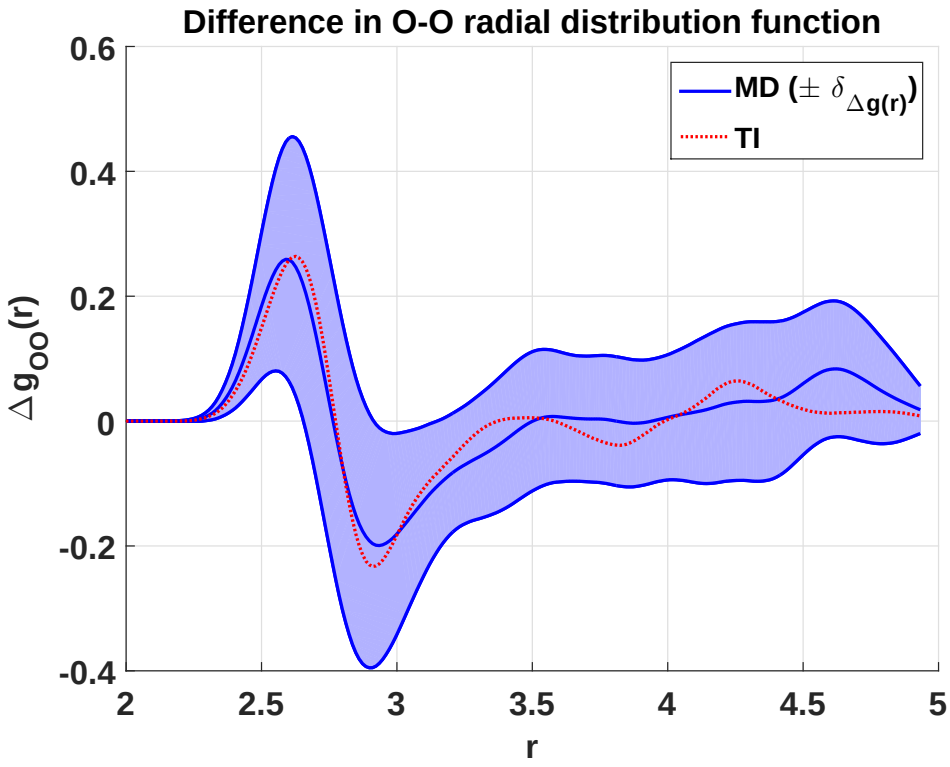
Here we considered the case of mixing energies from the same DFT functional, obtained with different basis sets. This is a much smaller perturbation than mixing the energies obtained with two distinct functionals, and therefore should yield more accurate predictions. Differences in the energies and forces with respect to the initial function V_0 for this perturbation are smaller that they would be in the case of two distinct functionals, and therefore similar geometries in AIMD trajectories are expected. The more representative the geometry distribution is of a functional and parameters, the more realistic the energies of the snapshots to be used in the TI estimation will be.

For this test, we used the same 200 snapshots that were used in the TI pressure estimation test with DFT functionals that was discussedin the previous subscetion. Energies calculated with the vdW-DF-cx functional and the $^{(p)}$dζ+p basis set were represented by V_0 in the TI method, and those calculated with the same functional and the $^{(s)}$qζ+dp basis set were represented by V_1. All parameters (other than the basis set) used for energy calculations here were the same as those used in the full AIMD simulation (see previous subsection). The change in the oxygen-oxygen RDF $\Delta g_{OO}(r)$ introduced by changing the basis set from $^{(p)}$dζ+p to $^{(s)}$qζ+dp was then estimated with Eq. 5.4.

The TI estimated difference $\Delta g_{OO}(r)$ between the RDF prediced with the $^{(s)}$qζ + dp and $^{(p)}$dζ + p basis sets is plotted in Fig. 5.4, along with $\Delta g_{OO}(r)$ obtained from full AIMD trajectories. In addition, positions $r_{OO}^{1^{st}max}$ of the first maximum and r_{OO}^{min} of the first minimum, and the heights $g_{OO}^{1^{st}max}$ of the first peak and g_{OO}^{min} of the first valley of $g_{OO}(r)$, estimated with TI and from AIMD simulations for both basis sets are given in Table. 5.3.

The results in Fig. 5.4 and Table 5.3 demonstrate that both TI and AIMD predict a shift of the first peak to a smaller values of r with the $^{(s)}$qζ + dp basis set. AIMD predicts a shift of 0.02 Å in the first peak, while TI predicts a shift 0.03 Å. Taking into account statistical errors, the TI method accurately predicts the magnitude of this shift in the first peak.

In addition, a comparison of results reveals that the TI method is able to accurately reproduce the increase in the height of the first peak of $g_{OO}(r)$ with the $^{(s)}$qζ+dp basis set that is seen in the AIMD results. The AIMD simulations result in



(a)

Fig. 5.4 The difference between the $g_{OO}(r)$ calculated from an AIMD trajectory generated at 300 K with the $^{(s)}q\zeta$+dp basis and the one calculated with the $^{(p)}d\zeta$+p basis plotted along with the difference estimated with the TI method, using 200 snapshots from the $^{(p)}d\zeta$+p simulation. All calculations used the vdW-DF-cx functional, a 200 Ry mesh cutoff, 32 molecules per unit cell, and a 5×10^{-4} Bohr^{-3} DM tolerance.

a 0.11 increase in $g_{OO}^{1^{st}max}$, with statistical errors on the order of 0.1, while TI predicts a 0.09 increase.

The location of the first minimum in $g_{OO}(r)$, however, is not as well-defined, and contradictory results for this value show how statistical fluctuations can mask trends. The trajectory of the AIMD simulation predicts a decrease in g_{OO}^{min} and a shift of its location to larger values of r, while TI predicts a shift to smaller values of r. However, taking statistical errors into account, the TI estimates for both g_{OO}^{min} and r_{OO}^{min} are considered to be within the range predicted by AIMD.

These results demonstrate that the TI method can be applied to accurately predict changes in $g_{OO}(r)$ upon an improvement in basis set. This can be applied, for

Table 5.3 Positions $r_{OO}^{1^{st}max}$ of the first maximum and r_{OO}^{min} of the first minimum, and the heights $g_{OO}^{1^{st}max}$ of the first peak and g_{OO}^{min} of the first valley of $g_{OO}(r)$, obtained with the following methods: (i) an AIMD simulation using the $^{(p)}d\zeta$+p basis set, (ii) an AIMD simulation using the $^{(s)}q\zeta$+dp basis set, and (iii) the TI estimation of the $^{(s)}q\zeta$+dp RDF, using 200 snapshots from the $^{(p)}d\zeta$+p simulation. All calculations were performed with the vdW-DF-cx functional, a 200 Ry mesh cutoff, 32 molecules per unit cell, and a 5×10^{-4} Bohr^{-3} DM tolerance. The target temperature was 300 K for both AIMD simulations. All distances are given in Å.

Method	$r_{OO}^{1^{st}max}$	r_{OO}^{min}	$g_{OO}^{1^{st}max}$	g_{OO}^{min}
AIMD ($^{(p)}$dζ+p)	2.76 ± 0.01	3.30 ± 0.02	3.08 ± 0.10	0.48 ± 0.10
AIMD ($^{(s)}$qζ+dp)	2.74 ± 0.02	3.33 ± 0.05	3.19 ± 0.22	0.43 ± 0.13
TI ($^{(s)}$qζ+dp)	2.73	3.25	3.17	0.46

example, to basis convergence tests, or to estimating changes in the liquid water structure with a highly converged basis set after running a full simulation with a cheaper, less accurate basis set.

5.2.3 Application of the TI method to convergence tests

After completing the tests of the TI method presented in the previous section, we applied it to the estimation of the properties that were calculated in the convergence tests presented in Section 5.1, with the exception of the test involving the number of molecules per unit cell (Fig. 5.5). The TI method cannot be applied to the test of convergence with number of molecules per unit cell, because the values of the hybrid potential need to be evaluated for the same snapshots. The values of $\langle P \rangle$, $\langle \zeta \rangle$, $g_{OO}^{1^{st}max}$, and $g_{OO}^{1^{st}min}$ estimated with the TI method are plotted along with the values from full AIMD simulations (Table 5.2) for each basis set (Fig. 5.6), mesh cutoff (Fig. 5.7), and DM tolerance (Fig. 5.8).

A comparison of the AIMD results with the TI method reveals that in some cases the convergence of a property can be concealed by statistical fluctuations. For example, resulting values of $\langle P \rangle$ for different mesh cutoffs (Fig. 5.7a) and DM tolerances (Fig. 5.8a) from AIMD simulations suggest that the mesh cutoff might not be converged, and that a DM tolerance of at least 0.5×10^{-4} Bohr^{-1} may be required for convergence. The TI method, however, shows a clear convergence for 300 Ry

and a DM tolerance as high as 2×10^{-4} Bohr^{-1}. For all parameters, the TI method shows a faster convergence than suggested by the AIMD results.

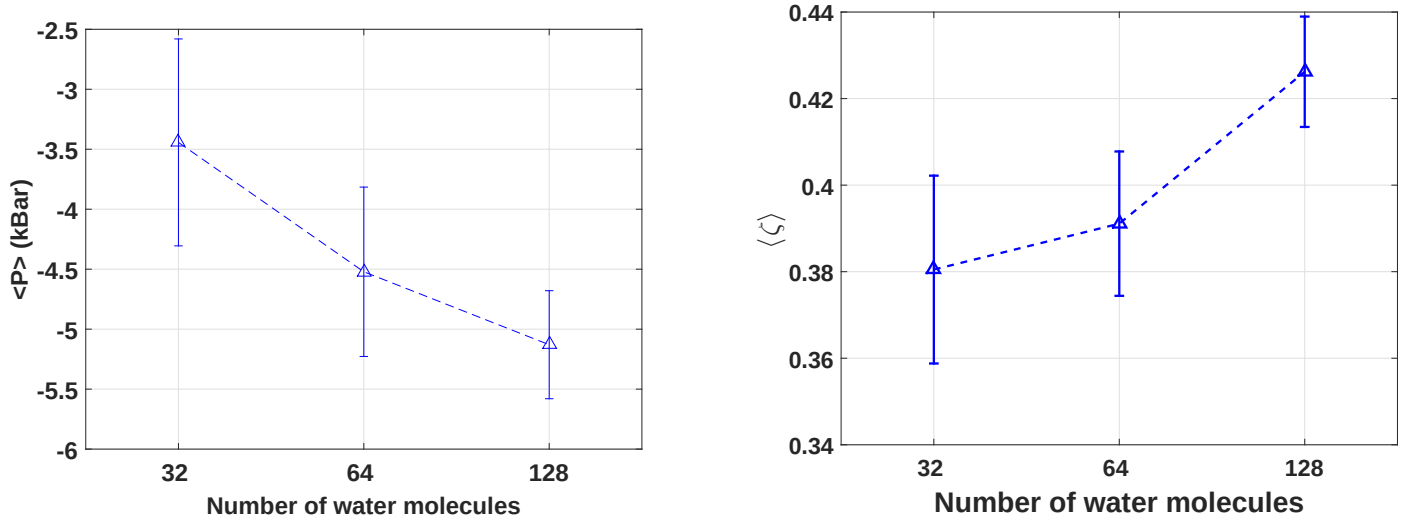
Taking into consideration statistical errors, the AIMD results suggest that $\langle P\rangle$, $\langle\zeta\rangle$, and position of the first maximum $g_{OO}^{1^{st}max}$ and the first minimum $g_{OO}^{1^{st}min}$ of the oxygen-oxygen RDF are converged within statistical errors for a system of 64 molecules, a $^{(s)}$qζ+dp basis set, a 300 Ry mesh cutoff, and a 0.5×10^{-4} Bohr^{-3} DM tolerance.

In the test of the basis sets, the TI method points to the tζ+dp basis set being better converged than suggested by the AIMD results. It does indicate, however, that the $^{(s)}$qζ+dp is better converged. The TI method also indicates a clear convergence with a 300 Ry mesh cutoff for all parameters, which is not as obvious from the AIMD results. The TI results also suggest that an even higher DM tolerance could be used.

In order to resolve doubts caused by statistical fluctuations in AIMD simulations, these simulations could be carried out for a much longer time, in order to obtain more data and thereby achieve a more accurate estimation of the errors. Ideally, the AIMD simulations should be carried out for as long as possible to reduce the statistical error. Unfortunately, AIMD simulations are too computationally expensive to keep them running for a sufficient amount of time. However, because the TI method only requires a one-time calculation with a small number of snapshots for each change in parameter to test for convergence, it has allowed us to resolve doubts about the convergence of some parameters in a much less computationally expensive manner.

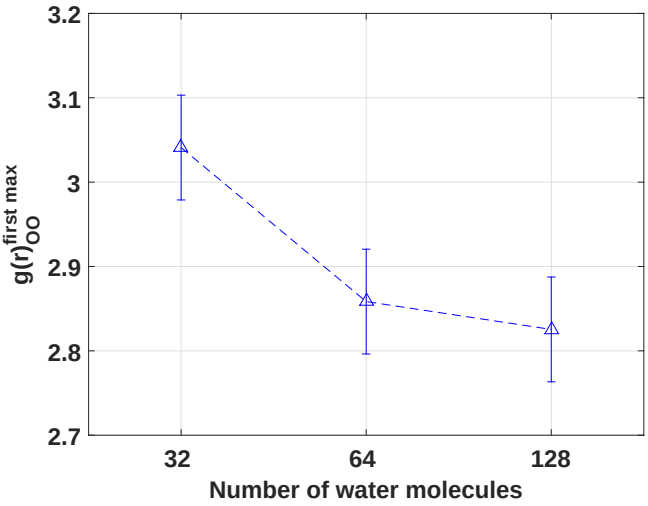
5.3 Conclusions

Several AIMD simulations were performed to test the convergence of the average simulation pressure $\langle P\rangle$, average order parameter $\langle\zeta\rangle$, and the first maximum $g_{OO}^{1^{st}max}$ and first minimum $g_{OO}^{1^{st}min}$ of the oxygen-oxygen radial distribution function $g_{OO}(r)$ with number of molecules per unit cell, basis set, mesh cutoff, and DM tolerance for liquid water. The dependence of these properties on average simulation temperature $\langle T\rangle$ was also studied, in order to adjust the property values and compare them all at the same effective temperature. However, because these properties are subject to large statistical fluctuations, the convergence of parameters can be difficult

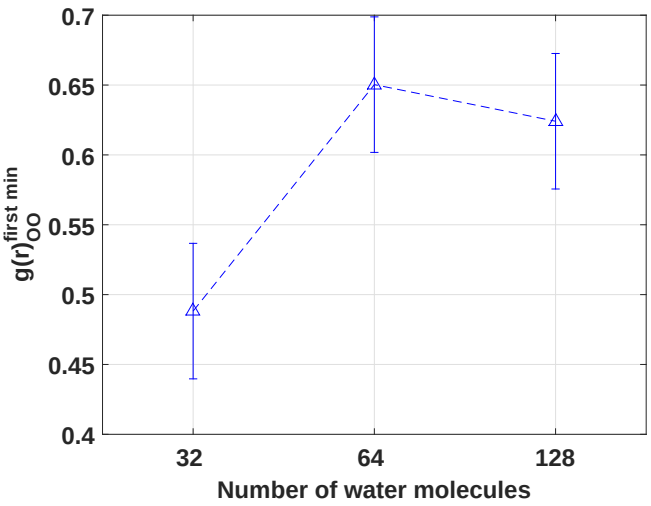


(a)

(b)



(c)



(d)

Fig. 5.5 Plots of $\langle P \rangle$ (a), $\langle \zeta \rangle$ (b), and the first maximum (c) and first minimum position (d) of $g_{OO}(r)$ for different system sizes: 32, 64, and 128 molecules per unit cell. All simulations are done with a ${}^{(p)}d\zeta$+p basis set, a 200 Ry mesh cutoff, and a DM tolerance of 1×10^{-4} Bohr^{-3}. Error bars are determined by Eq. 5.2

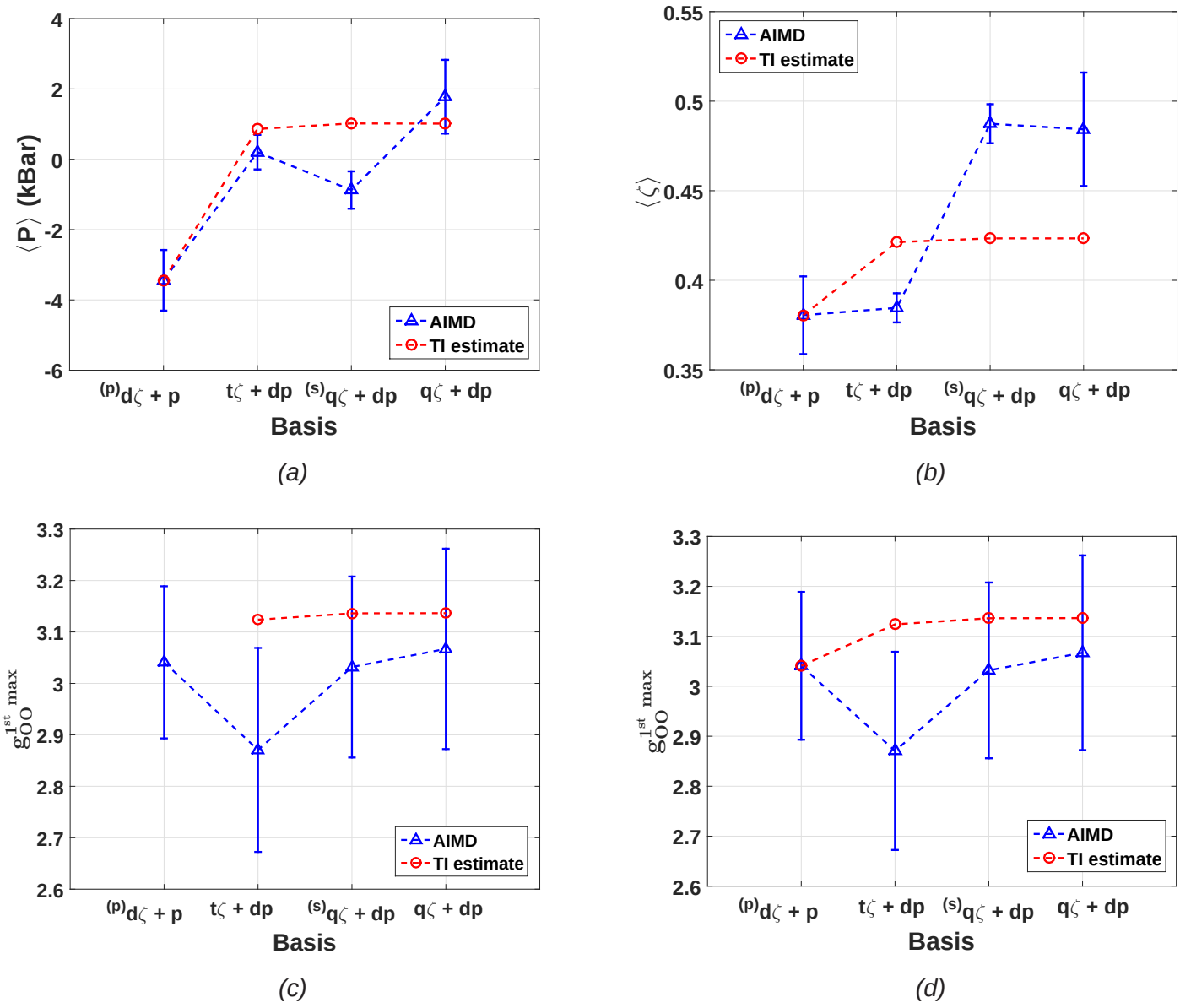


Fig. 5.6 Plots of $\langle P \rangle$ (a), $\langle \zeta \rangle$ (b), and the first maximum (c) and first minimum position (d) of $g_{OO}(r)$ for four different basis sets: ${}^{(p)}d\zeta$+p, tζ+dp, ${}^{(s)}q\zeta$+dp and qζ+dp. All simulations are done with 32 water molecules per unit cell, a 200 Ry mesh cutoff, and a DM tolerance of 1×10^{-4} Bohr^{-3}.

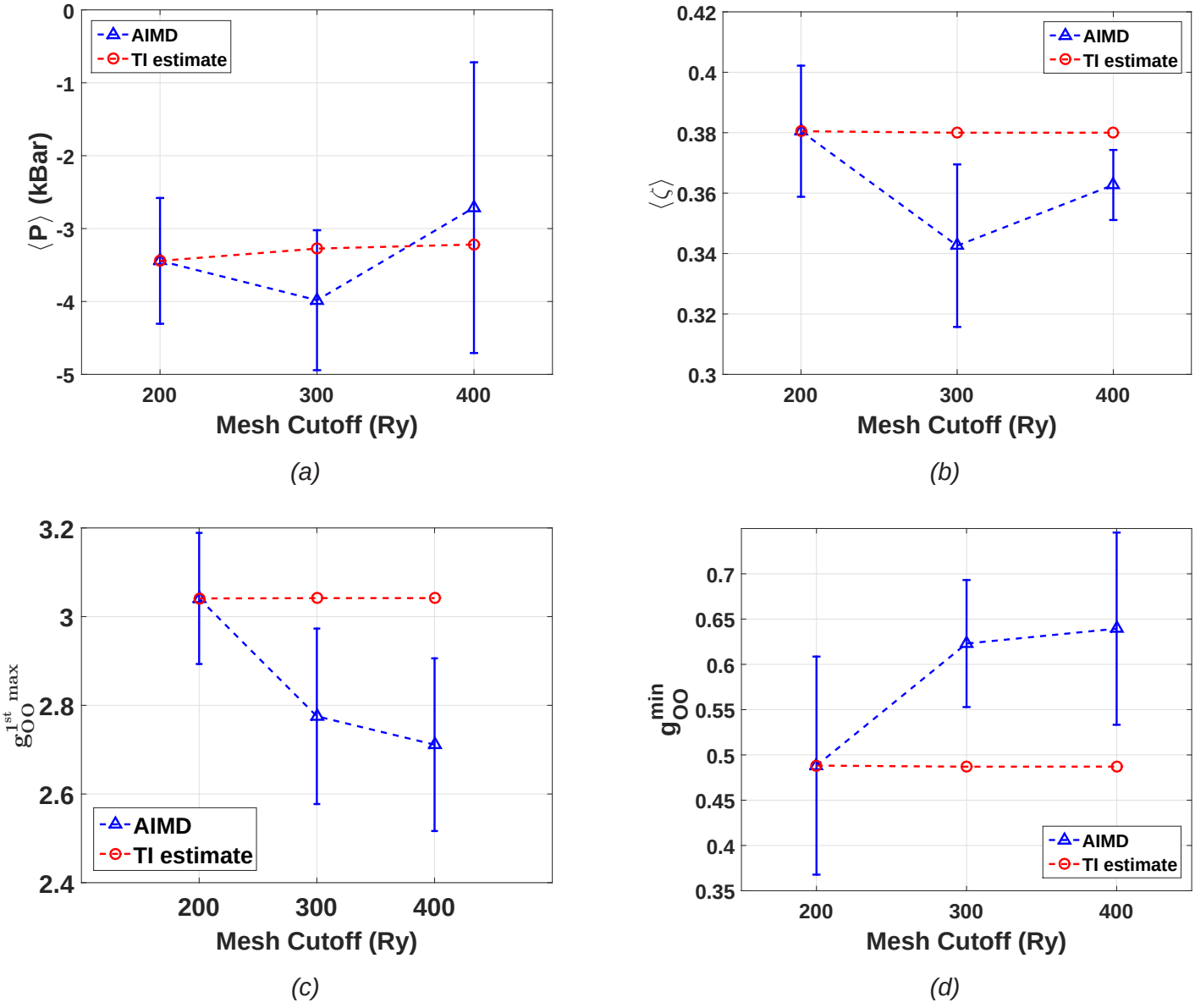


Fig. 5.7 Plots of $\langle P \rangle$ (a), $\langle \zeta \rangle$ (b), and the first maximum (c) and first minimum position (d) of $g_{OO}(r)$ for three different mesh cutoffs: 200 Ry, 300 Ry, and 400 Ry. All simulations are done with 32 water molecules per unit cell, a $^{(p)}d\zeta$+p basis set, and a DM tolerance of 1×10^{-4} Bohr^{-3}.

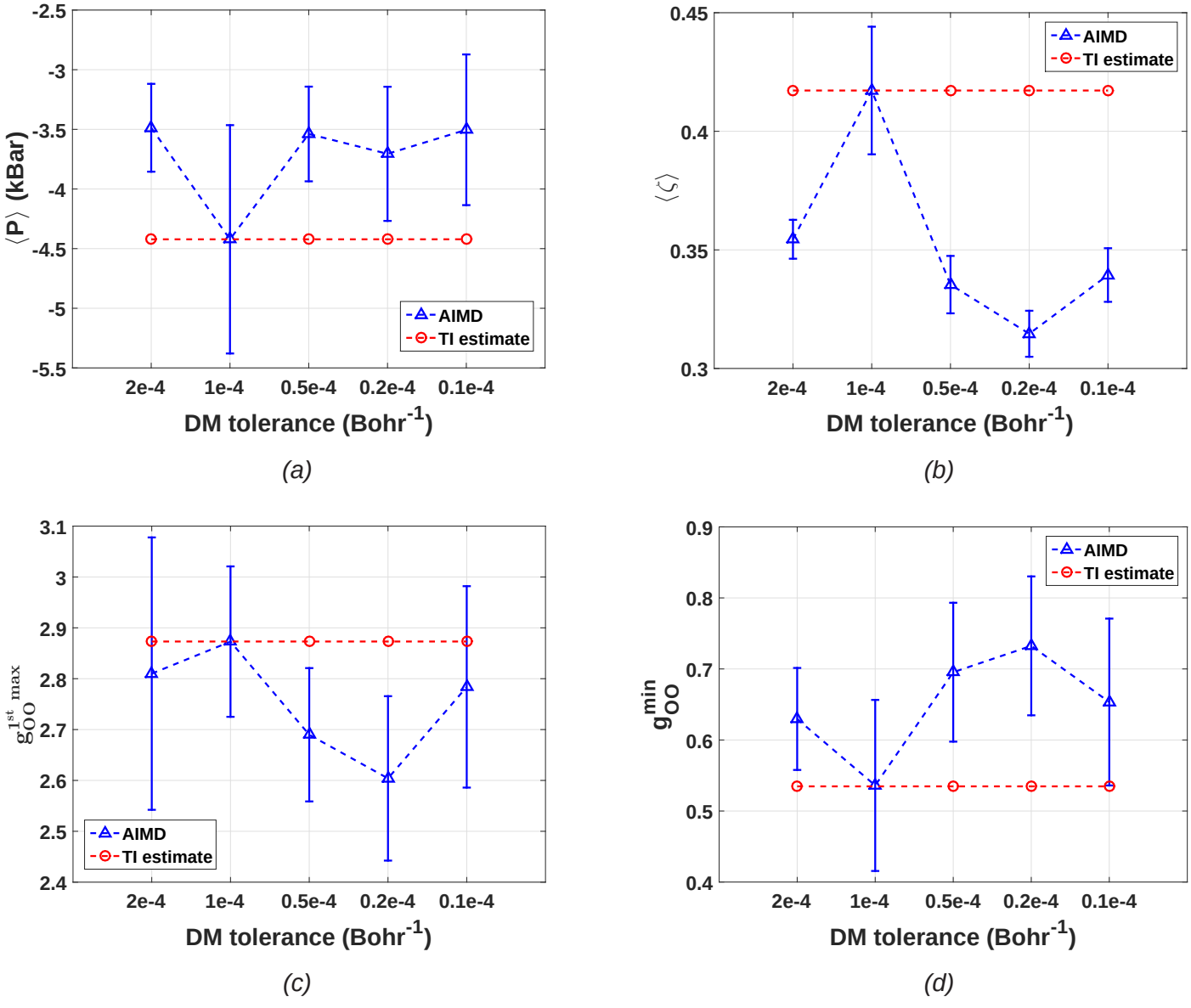


Fig. 5.8 Plots of $\langle P \rangle$ (a), $\langle \zeta \rangle$ (b), and the first maximum (c) and first minimum position (d) of $g_{OO}(r)$ for different DM tolerances (2×10^{-4}, 1×10^{-4}, 0.5×10^{-4}, 0.2×10^{-4}, and 0.1×10^{-4} Bohr^{-3}). All simulations are done with 32 water molecules per unit cell, a 200 Ry mesh cutoff, and a $^{(p)}d\zeta$+p basis set.

to determine unless sufficiently long simulations are performed. Unfortunately, this is far too computationally expensive.

In order to resolve doubts in convergence testing in a more efficient way, we have developed the Thermodynamic Integration (TI) method, named for its basis on ideas from thermodynamic integration. Using the TI method, only approximately 1% of the snapshots from a 20,000 step (10 ps) AIMD simulation trajectory are needed to estimate changes in certain properties with functional or parameter changes, making convergence testing significantly less computationally expensive.

In addition, because the same set of snapshots is used for each comparison in the TI method, it is subject to less statistical fluctuations than separate AIMD simulations. Therefore, this method is very useful is revealing convergence with a parameter that can easily be hidden by statistical fluctuations in the AIMD results. In our case, the TI method confirmed the convergence of some properties that was also seen in the AIMD results. It also revealed convergence in some areas where the AIMD results remained unclear, even after repeated attempts.

Aside from convergence testing, the TI method could also be applied to estimating property changes for improved parameters after running a full AIMD simulation with less computationally expensive parameters. For example, an AIMD simulation could be performed with the ${}^{(p)}$dζ+p basis set, which allows very quick simulations but is not fully converged. Then, the changes in liquid properties with the better converged but much more expensive ${}^{(s)}$qζ+p basis set could be estimated with the TI method. We have shown that the TI method correctly predicts changes in $g_{OO}(r)$ for this example.

In addition, while the TI method works best for small changes, such as changing a simulation parameter for the same functional, it could also be used for rough comparisons of completely different functionals. We have shown that, using snapshots from the AIMD trajectory of one functional, it can approximately predict changes in pressure that occur with an entirely different functional.

Chapter 6

Application of optimized functionals to gas and condensed phases of water

In Chapter 4, the DPPS method was introduced and its application to the optimization of the vdW-DF-cx functional was explained. Several DPPS-optimized functionals for water were introduced, but the two that will be focused on in this chapter are (i) CUSTOM-2B, optimized for 2-body energies of dimers and (ii) CUSTOM-MB, optimized for 2- and 3-body energies of dimers and trimers, respectively. Monomer-corrected versions (-Δ_1) of these functionals were also presented. The errors of 1-, 2-, and 3-body energies for the optimization dataset geometries with respect to the reference energies, calculated with the MB-pol potential [6], were analyzed for each functional. This is sufficient to demonstrate that the DPPS method works as intended; it results in a functional that better reproduces reference energies, with remaining errors due to physical constraints and limited degrees of freedom in the parameters. However, it is important to also study the performance of these functionals with water systems involving geometries not included in the optimization datasets and with a larger number of molecules. This is to ensure that the datasets used in the fit were sufficient, and to determine how well each functional describes water in the gas and condensed phases. Also, because of the differences in 1-, 2-, and 3-body energy errors in each functional, the relative accuracy of each one is expected to change as the number of molecules in the system of interest increases.

In this chapter, a study of the performance of vdW-DF-cx and the improvements introduced with the CUSTOM functionals for gas phase water hexamer equilibrium structures and water clusters from snapshots of a liquid water *ab initio* Molecular Dynamics (AIMD) simulation are discussed in Section 6.1. In Section 6.2, an analysis of AIMD simulations performed with these functionals to test their ability to predict to structural and thermodynamic properties of liquid water is presented.

6.1 Total energy analysis of clusters

6.1.1 Relative binding energies of hexamers

The study of hexamers is convenient for functional comparisons involving water, because they have been previously thoroughly researched in many different studies [114, 23, 130, 50, 56], and their energetics are known precisely [9]. A lot of attention is paid to water hexamers, because they are the smallest clusters in which the monomers form a fully three-dimensional structure. With a smaller number of monomers, the most stable isomers have ring-like geometries. However, with six or more molecules, compact geometries are more stabilized by bonding networks, such that there are small energy differences between different isomer configurations. Therefore, hexamers can be used as a quick test model for hydrogen-bond networks observed in condensed phases.

We tested the performance of vdW-DF-cx, CUSTOM-2B-Δ_1, and CUSTOM-MB-Δ_1 in calculations of the binding energies of five different hexamer isomers, labelled as follows: prism, cage, book, bag, and cyclic. The orientation of the water molecules in each isomer is shown in Fig. 6.1. The atomic positions used for these hexamer calculations correspond to the equilibrium configurations calculated by Temelso et al. [122] using second-order Møller-Plesset perturbation theory (MP2) [82] with the aug-cc-pVDZ basis set. A $^{(s)}$qζ+dp basis set, 300 Ry mesh cutoff, a 20 Å lattice constant, and a 0.5×10^{-4} Bohr^{-3} DM tolerance were used in calculations with all DFT functionals. These are the values for which the thermodynamic and structural properties of liquid water were found to be converged in Chapter 5, and are used in all DFT functional tests on clusters referred to in this section.

Fig. 6.2 shows the relative binding energies of the different hexamer isomers calculated with all three DFT functionals, and these values are compared with those of

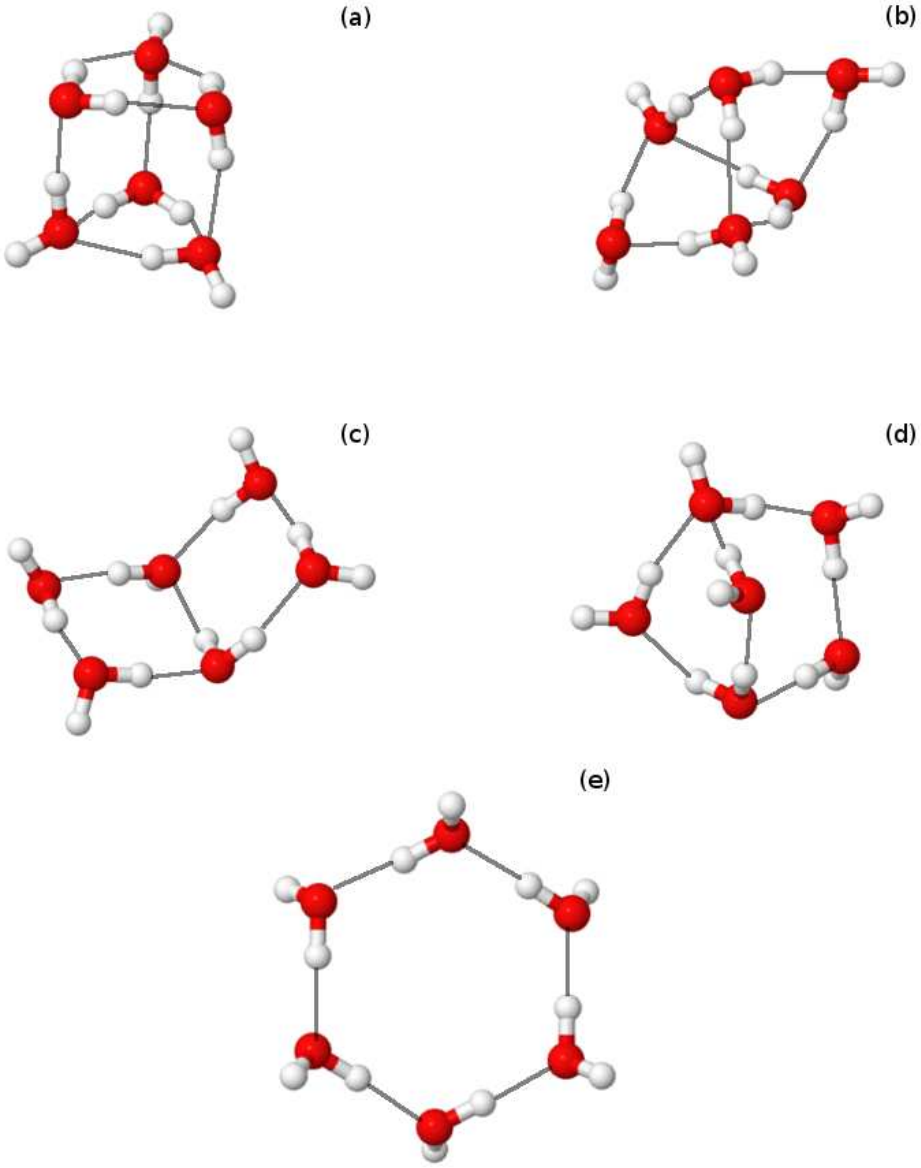


Fig. 6.1 The five isomers of the H_2O hexamer for which we analyze relative binding energies with different DFT functionals: (a) prism; (b) cage; (c) book; and (d) bag: and (e) cyclic. Red and gray spheres represent O and H atoms, with connecting lines showing hydrogen bonds.

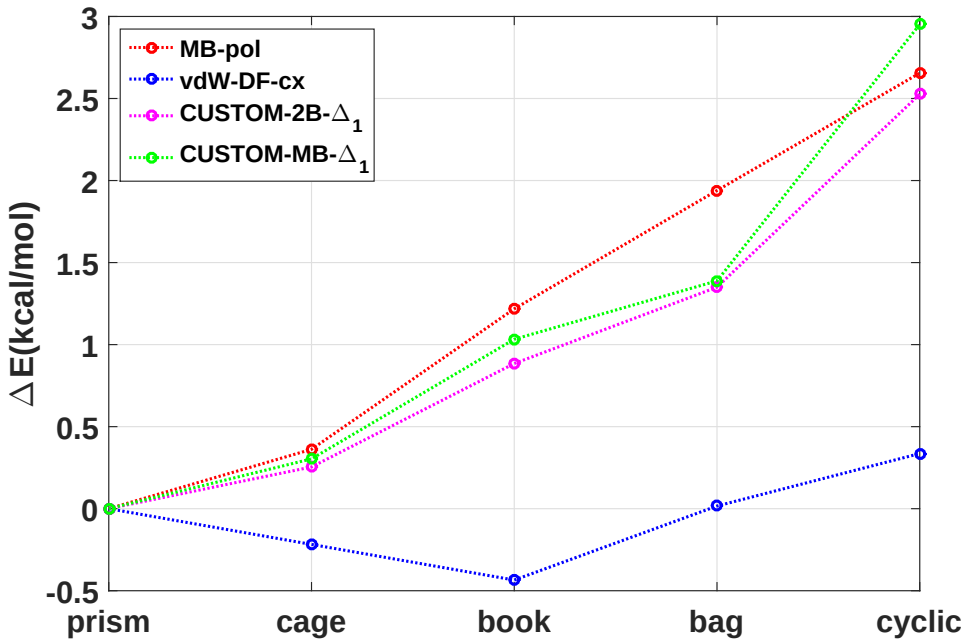


Fig. 6.2 Binding energies (in kcal/mol) of the prism, cage, book, bag, and cyclic water hexamer isomers shown in Fig. 6.1, relative to the binding energy of the prism isomer. The energies calculated for MP2 geometries are plotted, using MB-POL and three different DFT functionals: vdW-DF-cx, CUSTOM-2B-Δ_1, and CUSTOM-MB-Δ_1.

MB-pol. MB-pol was previously found to closely reproduce relative hexamer binding energies predicted at the CCSD(T) complete basis set (CBS) limit [78]. It correctly predicts that the order of stability of the hexamers, such that the prism is the most stable and the cyclic structure the least stable. However, the non-optimized DFT approximation, vdW-DF-cx, incorrectly predicts that the cage and book are the most stable configurations. It underestimates the relative binding energies of all of the structures with respect to the prism configuration. Both CUSTOM-2B-Δ_1 and CUSTOM-MB-Δ_1 show improvements from the vdW-DF-cx functional, by as much as 2.2 kcal/mol for CUSTOM-2B-Δ_1 and 2.6 kcal/mol for CUSTOM-MB-Δ_1, due to the reduced error of 2- and 3-body energies. In both cases, this causes the relative stability of the hexamers to be in agreement with the predictions of MB-pol. In most cases, CUSTOM-MB-Δ_1 yields relative energies closest to those of MB-pol, due to the inclusion of 3-body energies in the dataset of the DPPS fit of this functional.

6.1.2 Total energy analysis of water clusters from liquid snapshots

To include a wide range of relevant geometries, the accuracy of the vdW-DF-cx and monomer-corrected CUSTOM functionals in predicting the energies of six-molecule clusters taken from snapshots of a AIMD liquid water simulation was tested. The snapshots were taken from an AIMD simulation done at 300 K for 15 ps with the vdW-DF-cx functional, using the $^{(s)}$qζ+dp basis set and 32 molecules per unit cell. The clusters were chosen by randomly selecting a molecule, and then choosing the 5 nearest neighbors, within a cutoff of 6.5 Å, which is the cutoff for 2-body interactions in MB-POL [5]. Fig. 6.3 shows the mean absolute deviation (MAD) of the energies E_i with respect to MB-pol per molecule of 200 different water clusters chosen from different liquid water snapshots, calculated using the following equation:

$$MAD = \frac{1}{N_c}\sum_{i=1}^{N_c} | \left(E_i - E_i^{MB-pol}\right) - \langle E_i - E_i^{MB-pol}\rangle |, \tag{6.1}$$

where N_c is the number of configurations and $\langle E_i - E_i^{MB-pol}\rangle = (1/N_c)\sum_{i=1}^{N_c}\left(E_i - E_i^{MB-pol}\right)$ is the average energy difference over all of the configurations, which aligns the zero of energy with the MB-pol data.

Both CUSTOM functionals show an improvement over vdW-DF-cx in the cluster energy calculations, but the CUSTOM-2B functional results in a lower MAD than CUSTOM-MB. Even though CUSTOM-MB is including the additional improvement of 3-body energy terms, CUSTOM-2B shows a greater improvement in the 2-body energies and also results in an improvement in the 1-body energies (see Figs. 4.16 and 4.20 in Chapter 4). When compared to the results of the hexamers in equilibrium discussed in the previous section, the liquid snapshots include a wider range of monomers, dimer and trimer configurations. The better description of the 1-body and 2-body energies with CUSTOM-2B for the wider range of geometries included in the 200 clusters may be overcompensating for the additional 3-body energy improvements seen with CUSTOM-MB, causing CUSTOM-2B to more accurately describe total energies.

With the monomer energy corrections included, all functionals show a large improvement from their uncorrected versions. The reduction of the MAD by $\sim 52\%$ with vdW-DF-cx-Δ_1 upon the inclusion of monomer energy correction indicates that most of the error is contained in the 1-body errors.

However, the vdW-DF-cx-Δ_1 functional continues to give a large MAD compared to CUSTOM-2B-Δ_1 and CUSTOM-MB-Δ_1, which both improve by $\sim 70\%$ from their uncorrected versions. This is due to the remaining 2- and 3-body errors in vdW-DF-Δ_1, which are reduced in the other two functionals. The CUSTOM-2B-Δ_1 functional performs best out of all of the functionals, reducing the MAD to 0.072 mHartree/molecule, while CUSTOM-MB-Δ_1 reduces it to 0.088 mHartree/molecule. Even after including the monomer energy corrections, CUSTOM-2B-Δ_1 gives better overall total energy calculations of the hexamers than CUSTOM-MB-Δ_1, which is likely a result of a less accurate description of the 2-body energies with CUSTOM-MB-Δ_1. There is also the possibility that the dataset used to fix 3-body energies with DPPS was not sufficiently large, and that many 3-body energies remain unimproved with CUSTOM-MB-Δ_1.

The significance of the inclusion of the 1-body correction term in the total energy calculations is in agreement with a study performed by Morales *et al* [83] with various DFT functionals, which showed that the 1-body correction yielded a large improvement in the MAD with respect to Quantum Monte Carlo (QMC), especially in the case of non-hybrid functionals (Fig. 6.4). They used two pairs of functionals that differ based on inclusion of the Hartree-Fock exchange in one of each of the pairs: (i) PBE [98] and PBE0 [99, 2], and (ii) BLYP [10, 61] and B3LYP [11].

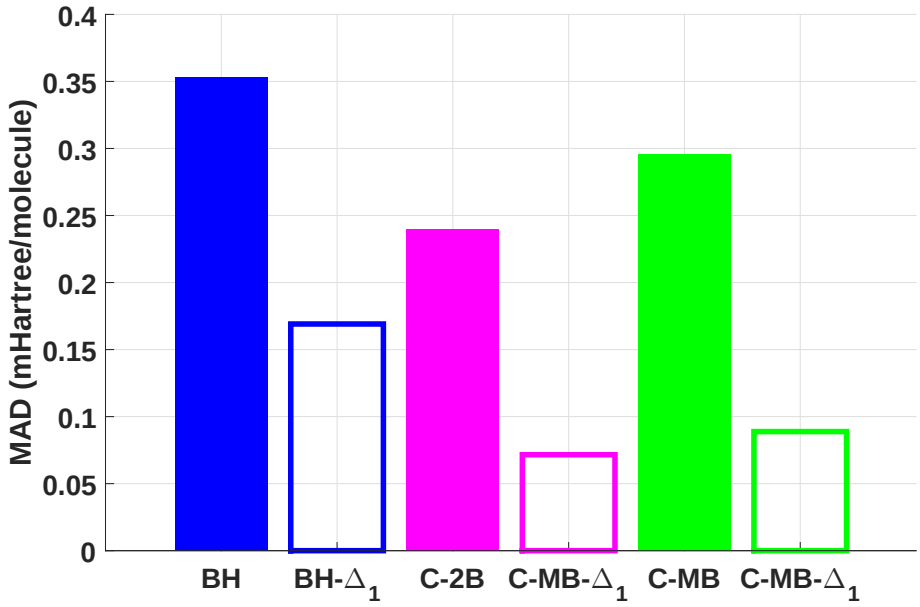


Fig. 6.3 MAD with respect to MB-pol of the total energy per molecule of 200 different 6-molecule water clusters taken from liquid water snapshots. The MAD was calculated with the following functionals: vdW-DF-cx (BH), vdW-DF-cx-Δ_1 (BH-Δ_1), CUSTOM-2B (C-2B), CUSTOM-2B-Δ_1 (C-2B-Δ_1), CUSTOM-MB (C-MB) and CUSTOM-MB-Δ_1 (C-MB-Δ_1). The liquid water snapshots were taken from a 15 ps AIMD simulation with the vdW-DF-cx functional at 300 K.

PBE0 and B3LYP are both hybrid functionals and therefore include a percentage of Hartree-Fock exchange, correcting some of the self-interaction error. Morales *et al* found that with the non-hybrid functionals, the largest contribution to the total energy MAD is the 1-body term. However, with the hybrid functionals, which include a reduction of self-interaction energy (SIE) error, there was not much difference between the uncorrected version and the 1-body corrected version.

We showed in Chapter 4 that significant SIE corrections are not possible by optimizing the GGA exchange alone. However, by adding the correction of the monomer energies, a majority of the error in the total energy due to SIE is eliminated. This is expected to show some improvements in the structure of liquid water predicted by AIMD simulations. Inclusion of monomer energy corrections results in a more realistic distribution of monomer geometries, shortening intramolecular O-H bonds, and therefore weakening the H-bond network and decreasing the structure

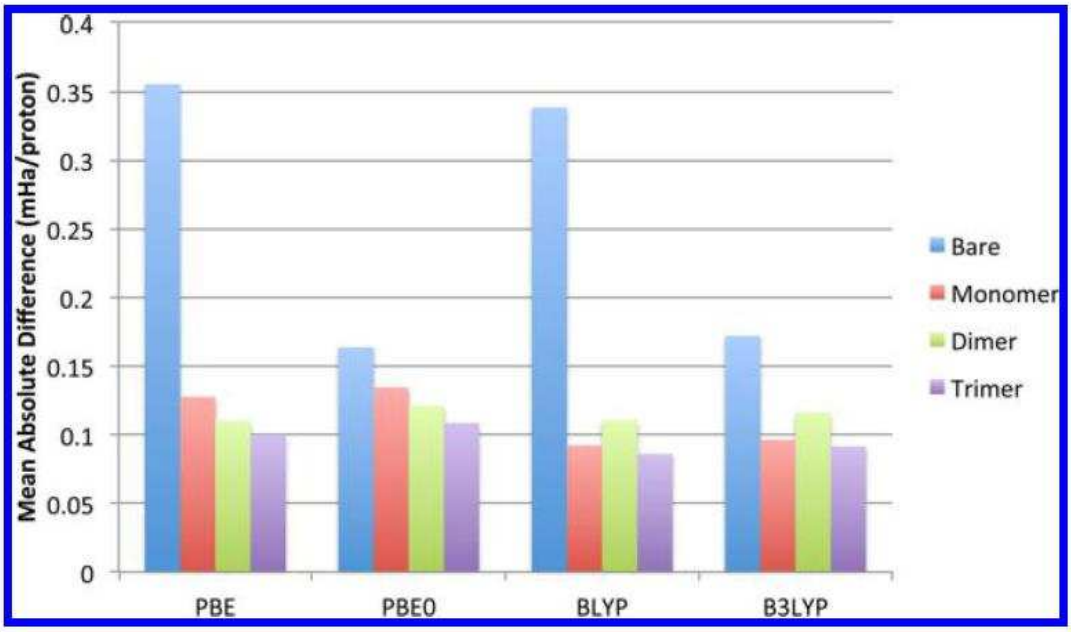


Fig. 6.4 This figure is taken from Ref. [83]. It plots the MAD with respect to QMC of 50 water snapshots from a vdW-DF2 [62] Path Integral Molecular Dynamics (PIMD) simulation, done at 300 K with 64 molecules per unit cell. Energies are calculated with 4 different DFT functionals with 1-body (red), 2-body (green), or 3-body (purple) energies corrected. Uncorrected energies are labelled Bare (blue).

of the liquid. Nonetheless, previous studies have shown mixed conclusions about whether or not the inclusion of SIE corrections improves the description of liquid water [46].

6.2 Ab initio Molecular Dynamics results

AIMD simulations of liquid water were performed with the vdW-DF-cx and CUSTOM functionals in order to get a better understanding of the importance of an accurate description of 1-, 2- and 3-body energies, considered both separately and simultaneously, in describing condensed phases of water. Simulations were done with vdW-DF-cx, CUSTOM-2B, CUSTOM-MB, and with all of their monomer-corrected versions ($-\Delta_1$). By including 1-body energy corrections, a non-self-consistent correction of the SIE is being considered, which has been suggested to be important for a proper description of the liquid water structure.

Table 6.1 Average values of the following properties obtained from AIMD simulations with different functionals: average pressure $\langle P \rangle$, average temperature $\langle T \rangle$, and average order parameter $\langle \zeta \rangle$. The values are averaged over the last 10 ps of each simulation. For reference, the value of $\langle \zeta \rangle$ obtained with the MB-pol simulation is 0.0210 nm. A $^{(s)}$dζ+p basis set, 300 Ry mesh cutoff, and a 0.5$\times 10^{-4}$ Bohr^{-3} DM tolerance were used for all AIMD simulations.

Functional	$\langle P \rangle$ (kBar)	$\langle T \rangle$ (K)	$\langle \zeta \rangle$ (nm)
vdW-DF-cx	-3.2 + 0.2	302 +/- 2	0.0464 +/- 0.005
vdW-DF-cx-Δ_1	-2.6 + 0.2	296 +/- 1	0.0388 +/- 0.003
CUSTOM-2B	-6.3 + 0.3	297 +/- 1	0.0359 +/- 0.003
CUSTOM-2B-Δ_1	-6.6 + 0.3	299 +/- 1	0.0319 +/- 0.005
CUSTOM-MB	-7.4 +/- 0.2	304 +/- 1	0.0252 +/- 0.004
CUSTOM-MB-Δ_1	-7.5 +/- 0.2	295 +/- 1	0.0229 +/- 0.002

For all AIMD simulations we used 128 molecules, a $^{(p)}$dζ+p basis set, a 300 Ry mesh cutoff, and a 0.5$\times 10^{-4}$ Bohr^{-3} DM tolerance. While better convergence was observed for a larger basis set in Chapter 4, we chose the $^{(p)}$dζ+p basis set in order to decrease the computational cost. With $^{(p)}$dζ+p, the simulations are up to $\sim$11 times faster than with $^{(s)}$qζ+dp. The errors introduced in structural and thermal properties by the basis can then be estimated using the TI method, which was discussed in Chapter 4. We also chose to use 128 water molecules per unit cell in order to obtain more statistical data in a shorter simulation time, in the AIMD simulations and an MD simulation done with the MB-pol potential.

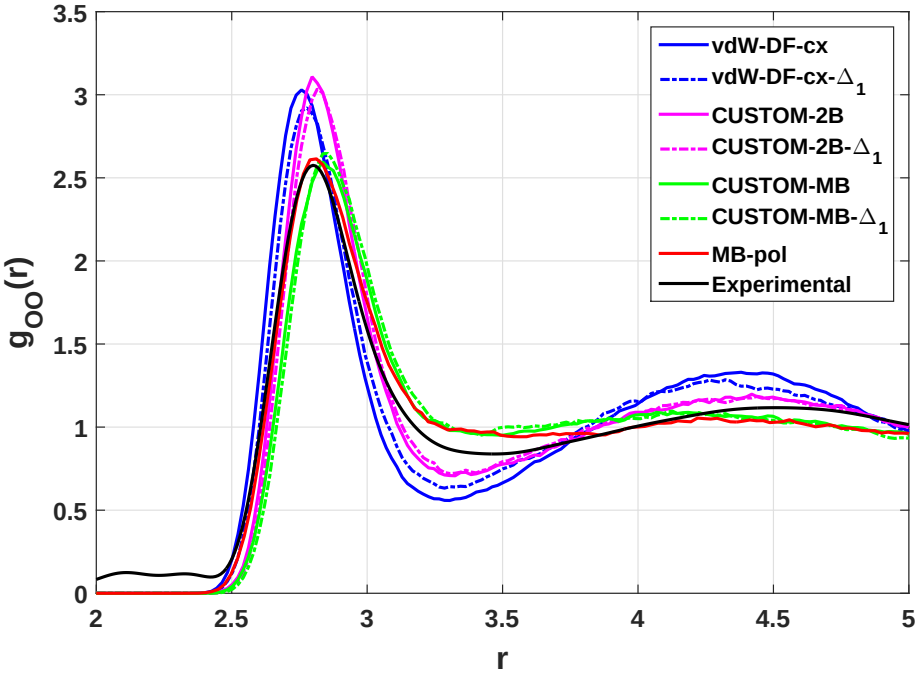
In all simulations, equilibration was accomplished with temperature annealing via velocity rescaling (NVT), and the production runs were performed using verlet integration [3] (NVE), with 0.5 fs time steps. The simulations were equilibriated at a temperature of 300 K during 5 ps, and after equilibration the simulations were continued for another 10 ps of production. They were performed at constant volume (fixed cell size and shape, under periodic boundary conditions), which was determined by the experimental equilibrium density of $\sim$1.0 g/cm^3 [1].

Values of the average simulation temperature $\langle T \rangle$, pressure $\langle P \rangle$, and order parameter $\langle \zeta \rangle$ obtained from the AIMD simulations with all six functionals are provided in Table 6.1. A $\langle P \rangle$ of $\sim$0.001 kBar would indicate an equilibrium density in agreement with experiments at ambient pressure. However, all six functionals result in a negative value for $\langle P \rangle$, indicating equilbrium densities that are too high. A change

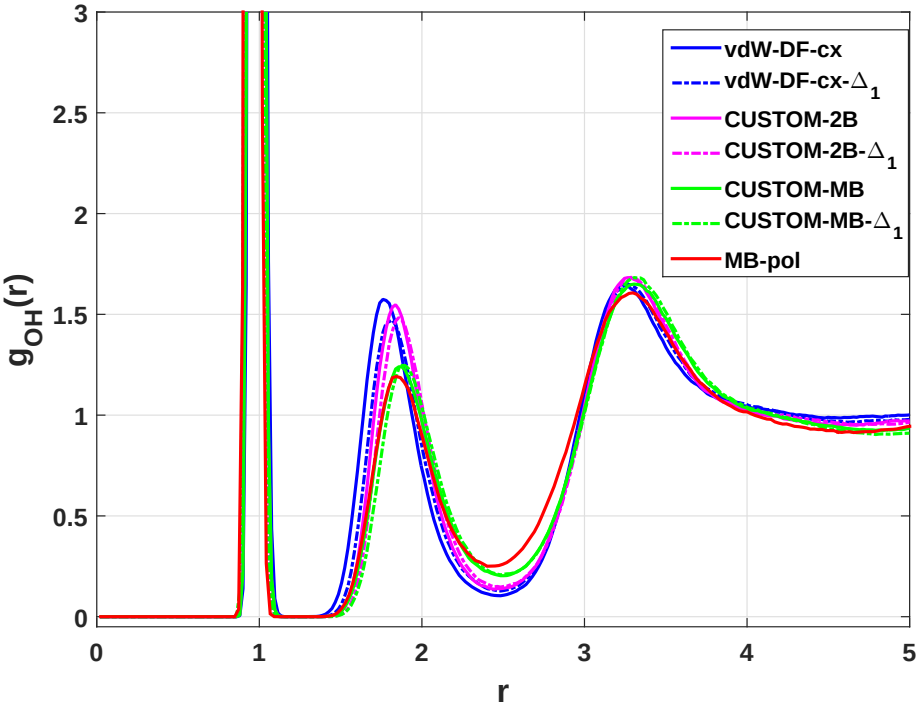
Table 6.2 RDF properties obtained from the same AIMD simulations described in Table 6.1: positions $r_{OO}^{1^{st}max}$ and $r_{OO}^{2^{nd}max}$ of the first and second maximum, position r_{OO}^{min} of the first minimum, the heights $g_{OO}^{1^{st}max}$ and $g_{OO}^{2^{nd}max}$ of the first and second peaks, and the height g_{OO}^{min} of the first valley of $g_{OO}(r)$ (Fig. 6.5a). The RDFs are obtained from the trajectories over 15 ps of simulation. The last row shows the TI estimate of $g_{OO}(r)$ with CUSTOM-MB-Δ_1 and a $^{(s)}q\zeta$+dp basis set, using 200 snapshots from the AIMD trajectory using CUSTOM-MB-Δ_1 and a $^{(p)}d\zeta$+p basis set.

Functional	$r_{OO}^{1^{st}max}$	$r_{OO}^{1^{st}min}$	$r_{OO}^{2^{nd}max}$	$g_{OO}^{1^{st}max}$	$g_{OO}^{1^{st}min}$	$g_{OO}^{2^{nd}max}$
MB-pol	2.81	3.47	4.22	2.57	0.914	1.05
vdW-DF-cx	2.76	3.31	4.38	3.03	0.558	1.33
vdW-DF-cx-Δ_1	2.78	3.29	4.32	2.92	0.633	1.29
CUSTOM-2B	2.80	3.32	4.42	3.11	0.707	1.20
CUSTOM-2B-Δ_1	2.82	3.31	4.44	3.04	0.721	1.19
CUSTOM-MB	2.84	3.50	4.07	2.60	0.951	1.05
CUSTOM-MB-Δ_1	2.85	3.42	4.13	2.65	0.952	1.05
CUSTOM-MB-Δ_1 ($^{(s)}$qζ+dp)	2.84	3.42	4.12	2.60	0.902	1.13

of basis from $^{(p)}$dζ+p to $^{(s)}$qζ+d causes an increase in $\langle P \rangle$ of about 4.5 kBar, as estimated by the TI method. However, even with a better basis, the CUSTOM functionals still give a negative $\langle P \rangle$ and an equilibrium density that is too large. This is expected, as we are not taking nuclear quantum effects into account in these simulations. The inclusion of quantum nuclear effects is expected to lower the density of the liquid [89].

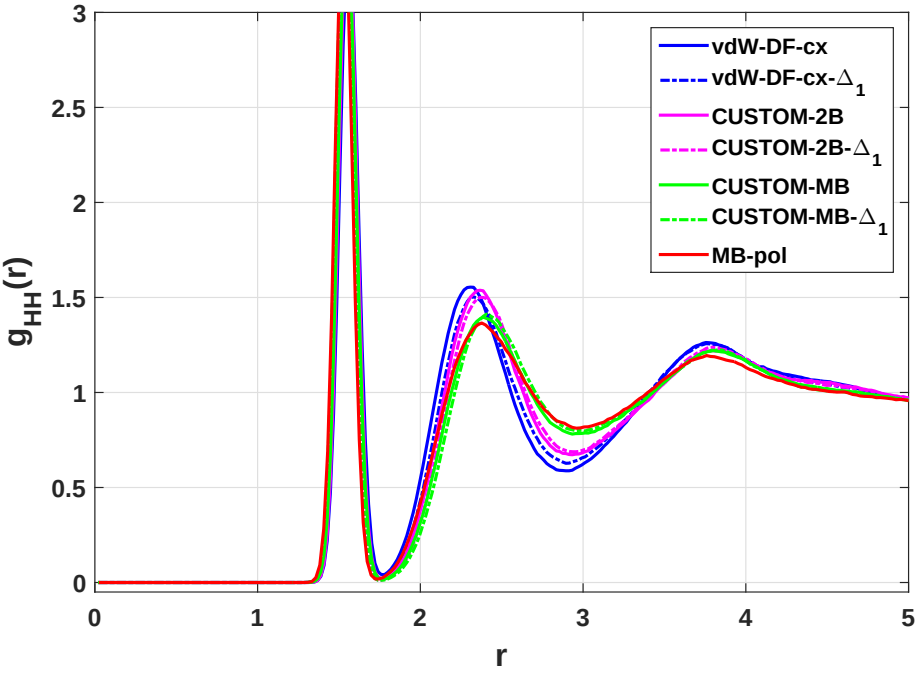

vdW-DF-cx
vdW-DF-cx-Δ1
CUSTOM-2B
CUSTOM-2B-Δ1
CUSTOM-MB
CUSTOM-MB-Δ1
MB-pol
Experimental
gOO(r)
r

(a)


vdW-DF-cx
vdW-DF-cx-Δ1
CUSTOM-2B
CUSTOM-2B-Δ1
CUSTOM-MB
CUSTOM-MB-Δ1
MB-pol
gOH(r)
r

(b)

Fig. 6.5



(c)

Fig. 6.5 Oxygen-oxygen RDF $g_{OO}(r)$ (a), oxygen-hydrogen RDF $g_{OH}(r)$ (b), and hydrogen-hydrogen RDF $g_{HH}(r)$ (c) resulting from DFT AIMD simulations with the following functionals: vdW-DF-cx (blue), vdW-DF-cx-Δ_1 (blue dash), CUSTOM-2B (magenta), CUSTOM-2B-Δ_1 (magenta dash), CUSTOM-MB (green), CUSTOM-MB-Δ_1 (green dash). For all DFT simulations, 128 molecules per unit cell, a ${}^{(p)}d\zeta$+p basis set, 300 Ry mesh cutoff, and 0.5×10^{-5} Bohr^{-1} DM tolerance were used. Also plotted for reference are the RDFs obtained from an MB-POL MD simulation (red) in Figures (a)-(c) and the experimental RDF $g_{OO}(r)$ (black) in (a) and for comparison. Data for the experimental RDFs was taken from Ref. [117].

The order parameter ζ was proposed by Russo and Tanaka [112] as a method of measuring the local translational order in the second shell of neighbors in liquid water. For a water molecule, ζ is defined as the difference between the distance to its first neighbor not considered hydrogen bonded $r^{1^{st}}_{non-H-bonded}$ and the distance to the last neighbor that is hydrogen bonded $r^{last}_{H-bonded}$. The value of $\langle\zeta\rangle$ obtained from the MB-pol simulation is 0.0213 nm. All DFT functionals yield values of $\langle\zeta\rangle$ that are too high in comparison (Table 6.1), indicating liquids that are too structured. The CUSTOM-MB and CUSTOM-MB-Δ_1 functionals are in closest agreement with

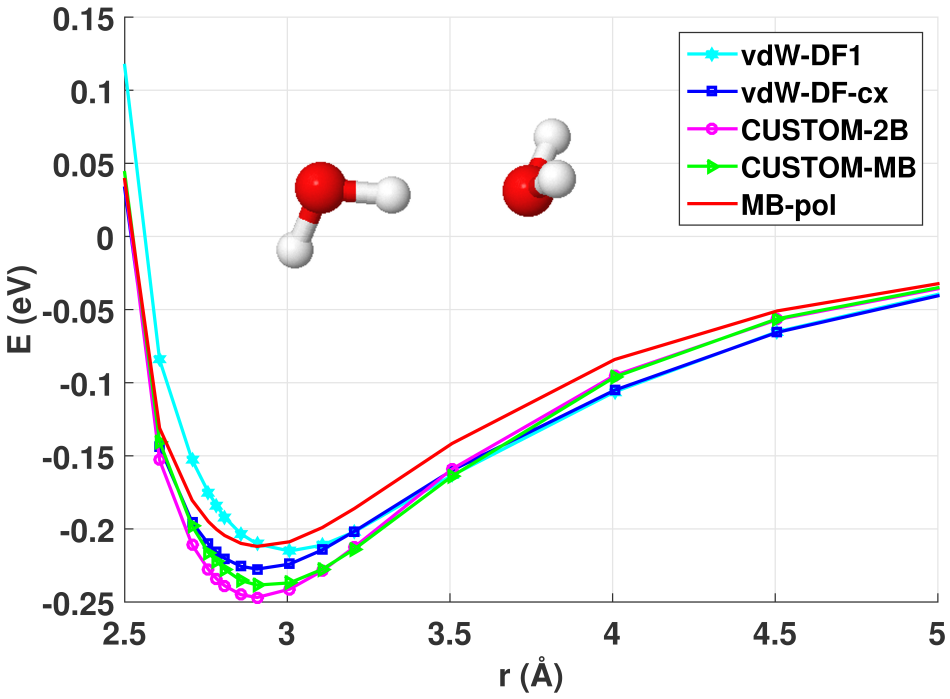
MB-pol in terms of second shell neighbor translation order, with values of $\langle\zeta\rangle$ equal to 0.0252 nm and 0.0229 nm, respectively.

As an important factor in the analysis of the structural properties of water determined by the DFT functionals and MB-pol, the oxygen-oxygen radial distribution function (RDF) $g_{OO}(r)$ (Fig. 6.5a), the oxygen-hydrogen RDF $g_{OH}(r)$ (Fig. 6.5b), and the hydrogen-hydrogen RDF $g_{HH}(r)$ (Fig. 6.5c) were calculated using the trajectories from the AIMD simulations performed. The positions $r_{OO}^{1^{st}max}$ and $r_{OO}^{2^{nd}max}$ of the first and second maximum, respectively, the position r_{OO}^{min} of the first minimum, as well as the heights of the first peak $g_{OO}^{1^{st}max}$, the first valley $g_{OO}^{1^{st}min}$ and the second max $g_{OO}^{2^{nd}max}$ of $g_{OO}(r)$ are all provided in Table 6.2.

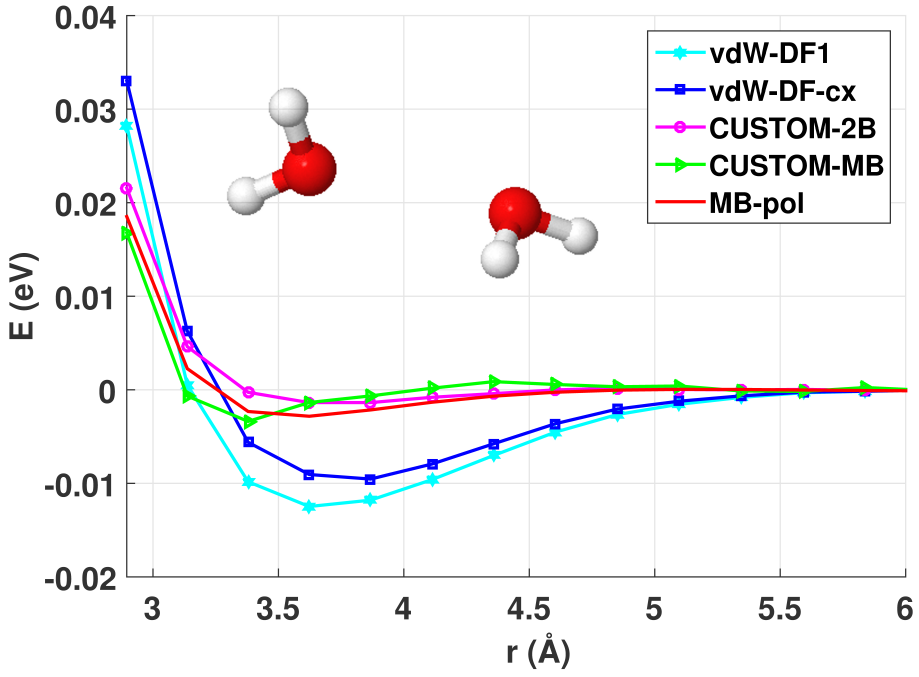
As an aid in the understanding the origin of the differences in $g_{OO}(r)$ as predicted by different functionals, the interaction energy curves for a dimer in an H-bonded configuration and for a dimer in a vdW-dominant configuration are plotted in Fig. 6.6. The monomer geometries in the H-bonded configuration are determined by the equilibrium geometry of the water dimer with MB-pol. For both configurations, the dimer is not allowed to relax, and the only degree of freedom is the O-O distance. The interaction energy curves compared are those of vdW-DF1, vdW-DF-cx, CUSTOM-2B, CUSTOM-MB and MB-pol.

The energy curves in Fig. 6.6 give us an idea of the relative strengths of the H-bond and vdW interactions with different functionals. Maintaining the proper balance between these two types of interactions is important. If the H-bond interactions are too strong relative to the vdW interactions, the H-bond network will be able to maintain the tetrahedral structure seen in ice, causing the liquid to be overstructured. This manifests as too many H-bonds being held in the first coordination shell, *i.e.*, a higher first peak in $g_{OO}(r)$. However, if the vdW interactions are too strong with respect to the H-bonds, too many molecules from the second coordination shell are pulled into the interstitial region by vdW forces, causing the second peak in $g_{OO}(r)$ to collapse. The second case has been shown to happen when the original vdW-DF1 by Dion *et al.* is used (Fig. 6.7) [129].

The vdW-DF-cx functional uses the same treatment of the non-local and local correlation as vdW-DF1, but it employs a different exchange energy functional. vdW-DF1 uses the revPBE exchange [135], while vdW-DF-cx uses LV-PW86R [14]. This alternative treatment of the exchange in vdW-DF-cx fixes the collapse of the second coordination shell of $g_{OO}(r)$ (Fig. 6.5a) that occurs with vdW-DF1 (Fig. 6.7). By examining the dimer energy curves (Fig. 6.6), it can be seen that the H-bonds


vdW-DF1
vdW-DF-cx
CUSTOM-2B
CUSTOM-MB
MB-pol
E (eV)
r (Å)
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
2.5
3
3.5
4
4.5
5

(a)


vdW-DF1
vdW-DF-cx
CUSTOM-2B
CUSTOM-MB
MB-pol
E (eV)
r (Å)
0.04
0.03
0.02
0.01
0
-0.01
-0.02
3
3.5
4
4.5
5
5.5
6

(b)

Fig. 6.6

Fig. 6.6 Interaction energy curves of two rigid water monomers as a function of the O-O distance (r) for an H-bonded-dominant configuration (6.6a) and a vdW-dominant configuration (6.6b). The dimer is not allowed to relax, and the only degree of freedom is r. In both graphs the energies have been shifted to have the zero at the largest separation. Orientations of the monomers are shown for each configuration. All energies were calculated with the $^{(p)}d\zeta$+p basis set.

are slightly stronger and the vdW interactions slightly weaker with vdW-DF-cx than with vdW-DF. This corrects the relative strength of these two types of interactions enough to resolve the collapse of the second coordination peak. However, vdW-DF-cx results in an over-structured liquid water, with the coordination peaks and the first minimum of $g_{OO}(r)$ being too pronounced in comparison to the result of MB-pol.

The H-bond interactions with vdW-DF-cx are stronger than with MB-pol, which is consistent with the increased height in the first coordination peak $g_{OO}^{1^{st}max}(r)$ from 2.57 to 3.03. The vdW interactions are also strengthened in comparison to MB-pol, but they are not strong enough to compete with the H-bonds and cause a collapse of the second coordination peak as seen with vdW-DF. However, the H-bonds continue to be too strong in comparison to the vdW interactions. This can be seen more clearly in Fig. 6.8, which shows the breakdown of $g_{OO}(r)$ into the H-bonds with the first coordination shell, the H-bonds with the second shell, and the vdW bonds plus other bonds. It is clear that vdW-DF-cx increases the populations of the first and second coordination shell, leaving the interstitial region depleted.

Effects of 2-body energy corrections on the radial distribution functions

In order to study the changes in the liquid water structure upon improving the 2-body energies, the RDFs obtained with the CUSTOM-2B functional are also plotted in Fig. 6.5a for comparison. With CUSTOM-2B, the first and second coordination peaks of $g_{OO}(r)$ are shifted to larger values of r with respect to the locations of the peaks of vdW-DF-cx. This brings the first peak location into better agreement with the MB-pol result (Table 6.2).

CUSTOM-2B also predicts a first coordination peak in $g_{OO}(r)$ that is higher than the one predicted by vdW-DF-cx. The significant differences between vdW-DF-cx and CUSTOM-2B are in the 2-body energies, and therefore in this case it is useful to consider the dimer interaction energy curves to get a better understanding of the

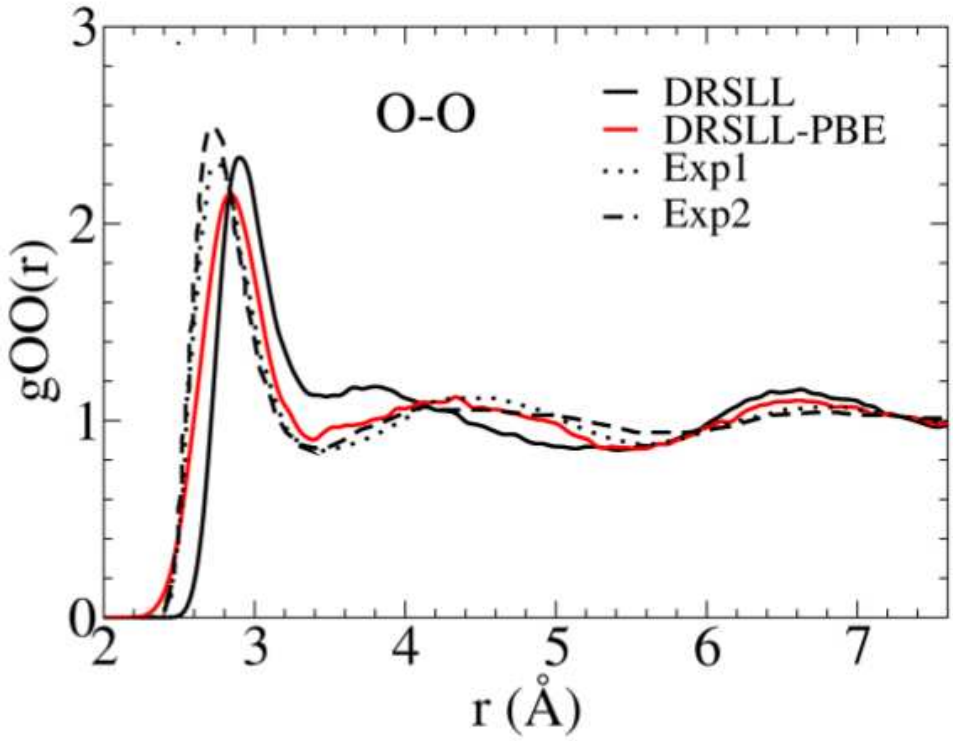


Fig. 6.7 Taken from Ref. [129]. Plot of the O-O RDF $g_{OO}(r)$ obtained with vdW-DF1, referred to as DRSLL here (black line), to show the collapse of the second coordination shell with this functional. This was taken from a simulation of water with a density of $1.0\ g/cm^3$ at 300 K, with a $^{(p)}d\zeta$+p basis set and 64 molecules per unit cell. The vdW-DF1 (DRSLL) result is compared to that of vdW-DF1 with PBE exchange (DRSLL-PBE, red), and the results from the experimental data of Refs. [45] (dotted line) and [119] (dashed line).

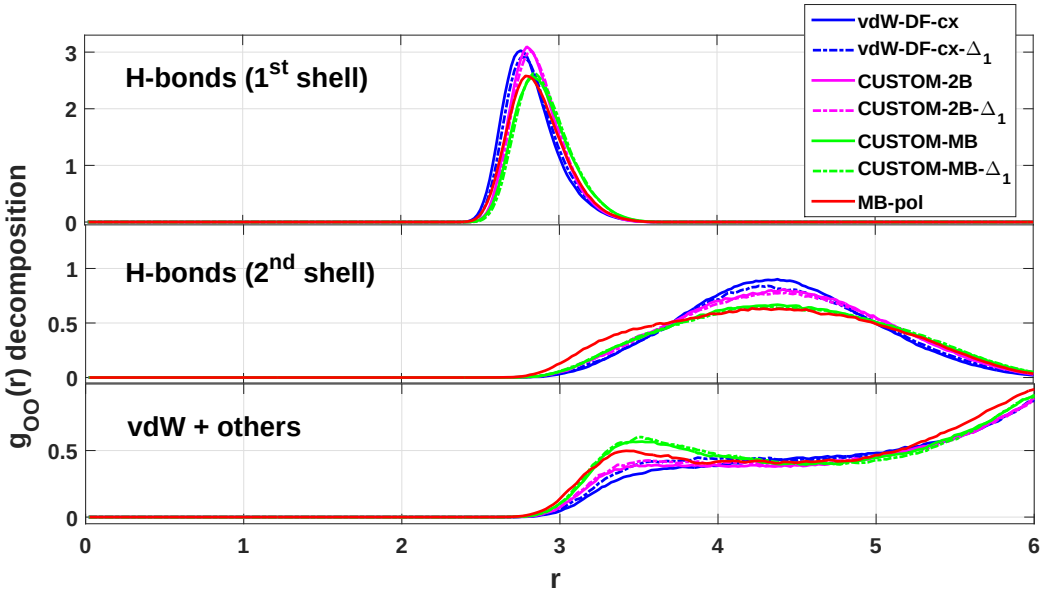


Fig. 6.8 Decomposition of $g_{OO}(r)$ from Fig. 6.5a. Top: first H-bonded neighbors, constituting the first coordination shell. Center: second H-bonded neighbors, constituting the second coordination shell. Bottom: remaining molecules which include vdW-bonded neighbors in the interstitial region and H-bonded molecules beyond the second coordination shell.

origin of changes seen in $g_{OO}(r)$. The increase in the height $g_{OO}^{1^{st}max}(r)$ of the first peak (from 3.03 to 3.11) can be understood as a result of stronger H-bonds and weaker vdW interactions between molecules, as demonstrated in the dimer energy curves of the H-bonded and vdW-dominant configurations of the dimer (Fig. 6.6). This encourages a higher population in the first coordination shell.

Despite the fact that the vdW interaction strength is weaker with CUSTOM-2B, the distribution of non-H-bonded configurations in the interstitial region does not decrease (Fig. 6.8). However, the second coordination peak of $g_{OO}(r)$ is lower, indicating that the bonds located around this separation r are weaker, allowing vdW interactions to shift the distribution towards the interstitial region. This brings CUSTOM-2B into better agreement with MB-POL in terms of the interstitial region and second coordination peak. CUSTOM-2B also results in improvements of $g_{OH}(r)$ (Fig. 6.5b) and $g_{HH}(r)$ (Fig. 6.5b), shifting the locations and heights to be in better agreement with MB-pol.

Effects of 3-body energy corrections on the radial distribution functions

Because CUSTOM-MB is optimized to both 2-body and 3-body energies of MB-POL, we can consider the change between the RDFs predicted with CUSTOM-MB and CUSTOM-2B to be mostly an effect of changes in the 3-body energies. Therefore, 3-body interactions play the largest role in the changes between these two functionals, and it is not very helpful to analyze the differences in their $g_{OO}(r)$ based on dimer interactions alone. CUSTOM-MB results in a less-pronounced first maximum in $g_{OO}(r)$ than CUSTOM-2B, decreasing it from 3.11 to 2.60 and bringing it into better agreement with MB-pol. It can be seen clearly in the plot of $g_{OO}(r)$ (Fig. 6.5a) and its decomposition (Fig. 6.8) that the distribution of H-bonds in the first and second shell decrease, and that the interstitial region becomes more populated by vdW-bonded configurations. With CUSTOM-MB, the position of the first peak of $g_{OO}(r)$ is shifted to a larger value of r (2.84 Å), so that the peak location is to the right of the one predicted by MB-pol. The second coordination peak is shifted to smaller values (r=4.13 Å), bringing it into closer agreement with MB-pol (Table 6.2). In both $g_{OH}(r)$ (Fig. 6.5b) and $g_{HH}(r)$ (Fig. 6.5b), CUSTOM-MB shifts the position of the the peaks and reduces their heights, bringing them closer to the results of MB-pol. Overall CUSTOM-MB and its 1-body-corrected version CUSTOM-MB-Δ_1 are in closest agreement with MB-pol out of all of the functionals we tested.

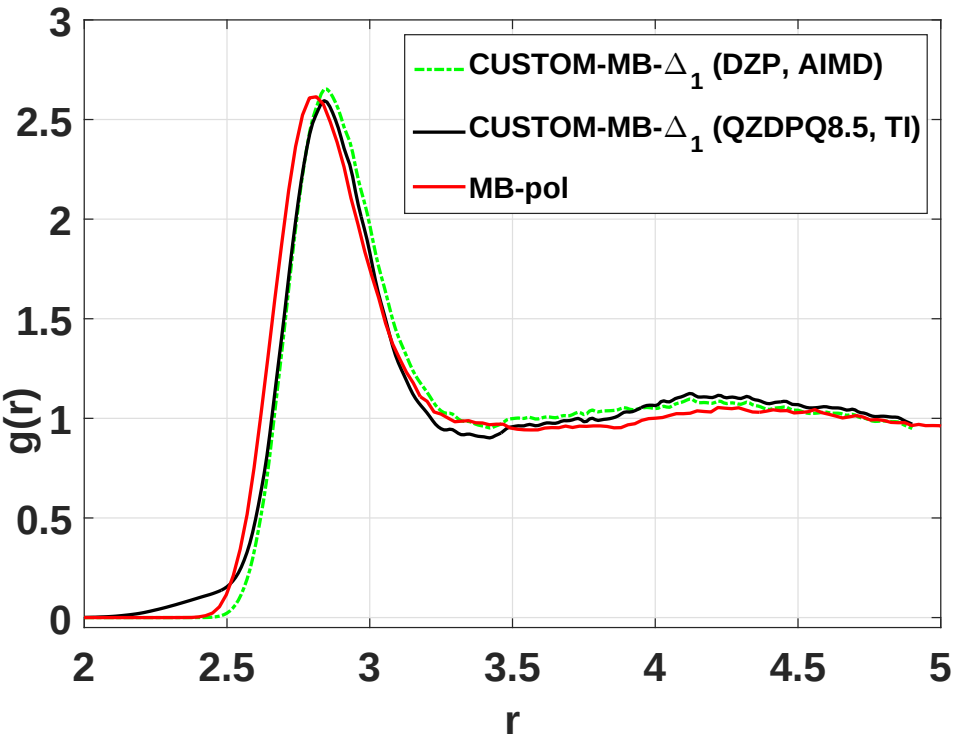
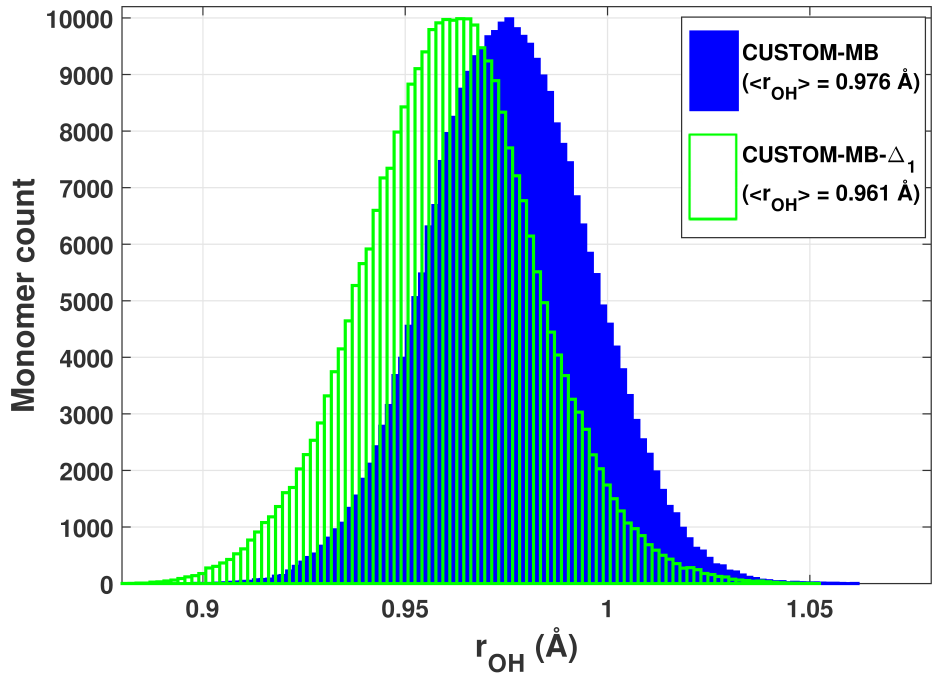
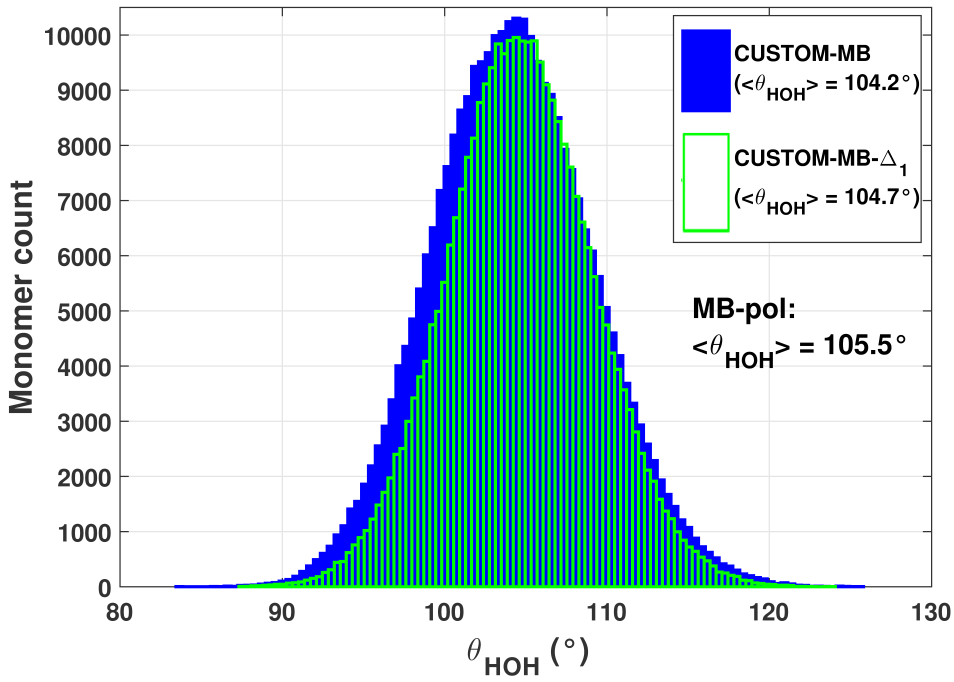


Fig. 6.9 TI estimation of oxygen-oxygen RDF $g_{OO}(r)$ with CUSTOM-MB-Δ_1 and the ${}^{(s)}q\zeta$+dp basis set (black), using 200 snapshots of the AIMD trajectory generated with CUSTOM-MB-Δ_1 and the ${}^{(p)}d\zeta$+p basis set, compared with the AIMD/ ${}^{(p)}d\zeta$+p (green, dot-dash) and MB-pol (red) results. All parameters other than the basis set are the same as those described in Fig. 6.5.

To test the effect of the basis set on the liquid structure predicted by CUSTOM-MB-Δ_1, the TI method was used to predict $g_{OO}(r)$ with the ${}^{(s)}$qζ+dp basis set, using 200 snapshots from the ${}^{(p)}$dζ+p trajectory. The TI-estimated $g_{OO}(r)$ is compared with the ${}^{(p)}$dζ+p and MB-pol simulation results in Fig. 6.9 and Table 6.2. The improvement in the basis set does not have a large effect on the structure, but it results in a slight reduction the height and location of the first peak in $g_{OO}(r)$, bringing it into slightly better agreement with MB-pol.


CUSTOM-MB
(<r_OH> = 0.976 Å)
CUSTOM-MB-Δ1
(<r_OH> = 0.961 Å)
Monomer count
r_OH (Å)
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0.9
0.95
1
1.05

(a)


CUSTOM-MB
(<θ_HOH> = 104.2°)
CUSTOM-MB-Δ1
(<θ_HOH> = 104.7°)
MB-pol:
<θ_HOH> = 105.5°
Monomer count
θ_HOH (°)
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
80
90
100
110
120
130

(b)

Fig. 6.10

Fig. 6.10 Histograms of the intramolecular distance between an oxygen atom and a hydrogen atom r_{OH} (6.10a) and the H-O-H angle for each molecule θ_{HOH} (6.10b). Averages are provided in the legend and are taken over 300,000 values. These values are for 10 molecules from each step of the 30,000 trajectories from the AIMD simulation with the CUSTOM-2B functional and the monomer corrected version CUSTOM-2B-Δ_1.

Effects of 1-body energy corrections on the radial distribution functions

Considering statistical fluctuations, there is not a significant change in the heights of the coordination peaks and interstitial region of the RDFs when 1-body energy corrections are added to the functionals. There is, however, a consistent shift to larger values of r in the first coordination peak that occurs upon the inclusion of 1-body corrections in each functional. With a better description of the monomer energies, the distribution of monomer geometries is in better agreement with those seen with MB-pol (Fig. 6.10). The average of the intramolecular distances r_{OH} between O and H atoms of the monomers of the AIMD trajectory produced with CUSTOM-MB is reduced from 0.976 Å to 0.961 Å (Fig. 6.10a), and the average intramolecular H-O-H angle θ_{HOH} is increased from 104.2° to 104.7° (Fig. 6.10b) when monomer corrections are included. This brings the geometry distribution closer to that of MB-pol, which results in $\langle r_{OH} \rangle = 0.958$ Å and $\langle \theta_{HOH} \rangle = 105.5°$. The inclusion of the 1-body energy corrections reduces the r_{OH} distances, leading to weaker hydrogen bond strength due to the anticorrelation effect. Weaker hydrogen bonds with molecules in the first coordination peak allow the molecules to shift farther apart, shifting the peak in $g_{OO}(r)$ to larger values of r.

6.2.1 Conclusions

In this chapter, the ability of the vdW-DF-cx, CUSTOM-2B, CUSTOM-MB functionals and their monomer-corrected versions to properly describe water systems ranging from the gas to liquid phase was considered. Results from MB-pol were used as references for comparisons.

First, to get an estimate the performance of each functional in water H-bond networks, the relative binding energies of five different water hexamers were calculated. It was shown that the vdW-DF-cx functional incorrectly predicts the cage

and book hexamers to be more stable than the prism, bag, and cyclic configurations. Both the CUSTOM-2B and CUSTOM-MB functionals were able to restore the proper order of stability in these hexamer isomers, with only a small improvement from CUSTOM-2B when CUSTOM-MB was used. This reveals that 2-body energy corrections applied to vdW-DF-cx are sufficient to restore the correct order in relative stability of these structures.

Both CUSTOM-2B and CUSTOM-3B were shown to reduce the error in the total energies of water cluster geometries found in snapshots of liquid water simulated with AIMD. A comparison of the results from each functional with their monomer corrected versions revealed that the largest contribution to the total energy error in all cases is the 1-body energy.

Nonetheless, we found that corrections of 1-body energies are not the most significant for obtaining a proper description of the structure of liquid water. Inclusion of the 1-body corrections in the functionals did not have as large of an effect on the level of structure seen in the liquid as 2- and 3-body corrections did. The 1-body corrections, however, did change the distribution of the intramolecular O-H distance, in turn weakening H-bonds due to the anticorrelation effect and shifting the first coordination peak to a larger value of r.

It was shown that a combination of 2-body and 3-body energy corrections is necessary to best predict the liquid water structure. The 2-body corrections included in the CUSTOM-2B functional do not significantly change the amount of structure in the liquid in comparison to vdW-DF-cx, but they significantly shift the first coordination peak to larger values of r, bringing it into better agreement with MB-pol. However, upon the additional inclusion of 3-body corrections in CUSTOM-MB, the liquid structure softens and looks more like the liquid predicted by MB-pol. The reduction of the over-structuring seen with the CUSTOM-MB functional demonstrates the importance of including 3-body energy corrections in water systems in the liquid phase.

Both CUSTOM functionals result in a liquid that has a density that is too high in ambient conditions. However, nuclear quantum effects, which were not included in the simulations, are expected to reduce the density. Nuclear quantum effects are important in calculations of liquid water, and should be included for an accurate analysis of a functional for this system. Studies using MB-pol with path-integral molecular dynamics (PIMD) have shown that, when nuclear quantum effects are taken into account, MB-pol gives a description of liquid water at ambient conditions

that is in close agreement with experiments for various structural, thermodynamic, and dynamical properties [76]. CUSTOM-MB-Δ_1 produces results for liquid water that are in close agreement MB-pol when nuclear quantum effects are not taken into consideration. Once they are taken into account, it is likely that this functional will also give results that are also in close agreement with experimental results.

Chapter 7

Conclusions

In this thesis, a new general method of optimizing energy functionals and force fields, called Data Projection onto Parameter Space (DPPS), has been presented and analyzed. We have tested this method on models within the framework of Kohn-Sham Density Functional Theory (DFT) implemented in the SIESTA package [118], and we have chosen water as the system of focus. Water is a system of great interest, not only because of its involvement in biological systems, but also because of its numerous anomalous properties. However, while force fields fitted to high accuracy data have been successful in accurately describing gas and condensed phases of water in recent years, DFT has remained behind in this area.

Most studies of DFT involve the development of exchange-correlation (XC) functionals, as the general form of this part of the total energy functional is not known exactly. One way to estimate the XC functional is to optimize it to a set of high accuracy data for a system of interest. In order to find the optimal exchange functional for water, we have applied the DPPS method, which avoids much of the guesswork involved in parametrization by projecting a vector of known data onto a vector of unknown parameters. The general DPPS method was presented in Chapter 4, along with its application to the optimization of a GGA exchange functional.

When developing new exchange-correlation functionals, the inclusion of physical constraints is crucial for the resulting functional to be physically feasible and transferable to a wide range of systems. With a limited dataset used in an optimization, functionals obtained with the DPPS method can have an impractical form that is not transferable to other systems or molecular configurations beyond the dataset used in the fit. Therefore, we have also developed a method of including physical constraints in the DPPS approach to optimization. By applying Bayes' Theorem of

probability, we include the physical constraints that are contained in pre-existing *ab initio* density functionals.

While DPPS is a general method that can be applied to many different function forms and types of data, in this thesis we have only presented its application to the optimization of GGA exchange functionals to sets of 1-, 2-, and 3-body energies of the many-body expansion for water. Energies were optimized to a reference dataset of those calculated with the MB-pol potential [4, 6], which represents water energies at the CCSD(T) [109] level of accuracy.

In a test of the DPPS method, we demonstrated that, by optimizing PBE exchange to a set of energies calculated with the PW86R functional, the form of PW86R exchange is accurately reproduced. This confirms that DPPS works as intended in the case of GGA exchange optimized to water energies. We then applied DPPS to the optimization of the GGA exchange part of the vdW-DF-cx functional for 1-body energies, 2-body energies, 3-body energies, and 2-body + 3-body energies simultaneously, with and without Bayesian constraints. The optimized functionals, both with and without constraints, were found to successfully reduce errors in the dataset energies. Although the constrained results showed less improvements in the energies, they provided more realistic functional forms. Considering the errors in each of the 1-, 2-, and 3-body energy terms, the functionals that were found to be most accurate were those fit to 2-body energies (CUSTOM-2B) and to both 2- and 3-body energies simultaneously (CUSTOM-MB).

Because of the limited degrees of freedom in the GGA exchange functional, we found it difficult to optimize this functional to 1-body energies simultaneously with other energy terms. Optimization to the 1-body energy term requires changes that cannot be reflected in the GGA exchange, resulting in unnatural changes in the functional form to compensate. These changes conflict with those needed for 2-body and 3-body energies corrections, making it impossible to obtain significant corrections to 1-body energies simultaneously with these terms through optimization of GGA exchange alone. Therefore, in order to adjust the 1-body energies calculated with the 2- and 3-body optimized functionals, a correction term was added to the total energy.

Before testing the DPPS-optimized functionals with *ab initio* Molecular Dynamics (AIMD) liquid water simulations, several simulations were performed to reveal which parameters produce converged results. In Chapter 5, results of AIMD simulations to test the convergence of the average simulation pressure $\langle P \rangle$, average order

parameter $\langle\zeta\rangle$, and the first maximum $g_{OO}^{1^{st}max}$ and first minimum $g_{OO}^{1^{st}min}$ of the oxygen-oxygen radial distribution function $g_{OO}(r)$ with number of molecules per unit cell, basis set, mesh cutoff, and DM tolerance for liquid water were presented. However, these properties obtained from AIMD simulation trajectories are subject to large statistical errors. Unfortunately, AIMD simulations are too computationally expensive to keep them running for a long enough time to sufficiently reduce the statistical fluctuations. As a result of the large statistical errors, the convergence of some parameters remained unclear with AIMD results alone. In order to resolve these doubts, we developed a more efficient method of convergence testing, based on ideas from thermodynamic integration, and therefore referred to as the TI method. With this method of convergence testing, calculations involving only 1% of the trajectory of an AIMD simulation is required, making it much more efficient. In addition, because the TI method uses the same trajectory to estimate changes for each simulation parameter, resulting property values with this method are subject to smaller statistical fluctuations than the AIMD results. After applying the TI method to the same convergence tests preformed with AIMD, we were able to resolve doubts about the convergence with some of the parameters tested and confirm convergence with others.

In addition to convergence testing, the TI method could be used to estimate system properties for very accurate but expensive parameters, after running a full AIMD simulation with cheaper, less accurate parameters for quicker results. We have also shown that the TI method can be used to obtain properties estimations for larger differences between methods, i.e., two completely different functionals rather than a simple parameter change. Thus, the TI method could also be applied to quick tests of new functionals in order to get an estimate of their accuracy for specific systems.

In Chapter 6, functionals optimized with Bayesian constrained DPPS were tested on gas and condensed phase water systems, and results were compared against those of MB-pol as a reference. In a test of the energies of water hexamer isomers, it was found that the introduction of 2-body energy corrections in CUSTOM-2B was sufficient for restoring the correct order of stability in the hexamers, with further improvements when the CUSTOM-MB functional was used. In addition, both optimized functionals were found to reduce the error in energy for a wider range of geometries in a study of random water clusters taken from AIMD

simulation trajectories. The results of these cluster calculations also revealed that the greatest contribution to the total energy error is contained in the 1-body term.

Considering the differences between the original vdW-DF-cx functional, and the DPPS-optimized versions, results from liquid water AIMD simulations suggest that the combination of 1-, 2-, and 3-body energy corrections is necessary to best predict the liquid water structure. While the 1-body energy corrections were not found to cause a significant change in the structuring of liquid water, they changed the distribution of monomer geometries, altering the intra-molecular O-H distances and H-O-H angles. This change in monomer geometry distribution was found to slightly shift the distribution of first neighbors to a larger intermolecular distance r due to weaker hydrogen bond strengths resulting from the anticorrelation effect. Similarly, the CUSTOM-2B-Δ_1 functional, which includes 2-body corrections, did not cause significant changes in the amount of structuring but did affect the location of H-bonded first neighbors, shifting the first peak of $g_{OO}(r)$ closer to that of MB-pol. The largest change in the structuring of the liquid, however, was seen with the additional inclusion of 3-body corrections in the CUSTOM-MB-Δ_1 functional, which softened the liquid water structure, bringing it into close agreement with the results of MB-pol.

While we were able to closely reproduce the results of MB-pol in Chapter 6 with the CUSTOM-MB-Δ_1 functional, it is important to consider nuclear quantum effects (NQE) that have not been included in the work presented in this thesis. It has been shown in previous studies [92, 46] that it is important to take NQE into account in order to accurately represent liquid water. When NQE quantum effects are taken into account, the MB-pol potential has been found to give a description of liquid water that is in close agreement with experiments for various structural, thermodynamic, and dynamical properties [76]. Because CUSTOM-MB-Δ_1, produces results for liquid water that are in close agreement with MB-pol without considering NQE, this functional should also give results that are also in close agreement with experimental results once these effects are taken into account. However, a more thorough study that incorporates NQE into the AIMD simulations should be carried out to confirm this.

The results of the cluster energies and liquid water AIMD simulations also reveal that, while the constrained DPPS method of optimizing the GGA exchange has provided a better description of water, there is still room for improvement in accuracy. One way to improve the resulting functionals would be to include a larger dataset of 2-body and 3-body energies in the DPPS optimization. However, even with a larger

dataset considered, the accuracy of the functional obtained is limited by its degrees of freedom. By optimizing the GGA exchange alone, the possible improvements in the functional are being restricted to those that can be reflected in the reduced density gradient s. Therefore, more accurate results could be obtained by optimizing an exchange functional with more degrees of freedom, or by including the optimization of the correlation functional. In addition, larger water systems or additional molecular systems could be included in the optimization dataset to obtain a more robust and general functional.

Chapter 7

Conclusiones

En esta tesis, un nuevo método general para optimizar funcionales de energía y campos de fuerzas, que se llama “Data Projection onto Parameter Space” (DPPS) ha sido presentado y analizado. Hemos comprobado este método con modelos dentro de la teoría de funcionales de densidad (DFT), implementado en el paquete de programación SIESTA, y hemos elegido el agua como el sistema de enfoque. El Agua es un sistema de gran interés, no solo por su participación en sistemas biológicos, sino también por sus propiedades anómalas. Sin embargo, mientras que los campos de fuerza de alta precisión han tenido éxito al describir fases gaseosas y condensadas del agua, la DFT ha quedado retrasada en este tema.

La mayoría de los estudios de DFT tienen que ver con el desarrollo de un funcional de intercambio y correlacion (XC), porque la forma general de esta parte del funcional de la energía total no es conocida de manera precisa. Una manera de estimar el funcional XC es optimizarlo a un conjunto de datos de alta precisión para un sistema de interés. Para encontrar el funcional de intercambio óptimo para el agua, hemos aplicado DPPS, lo que evita conjeturas (guesswork) en la parametrización, proyectando un vector de datos conocidos en un vector de parámetros desconocidos. El método general de DPPS fue presentado en el Capítulo 4, con su aplicación a la optimización de un funcional de intercambio GGA.

Cuando uno quiere desarrollar nuevos funcionales de densidad, la inclusión de constreñimientos es crucial para conseguir un funcional que sea físicamente factible y transferible a un rango grande de sistemas. Con un conjunto de datos limitado usado para la optimización, los funcionales obtenidos con DPPS pueden tener una forma que resulta poco práctica y que no es transferible a otros sistemas o a configuraciones moleculares fuera del conjunto de datos. Por lo tanto,

hemos desarrollado un método para incluir constreñimientos físicos en el método de DPPS. Aplicando el teorema de Bayes, incluimos los constreñimientos físicos que los funcionales de densidad preexistentes tienen en cuenta.

Mientras que DPPS es un método general que puede ser aplicado a muchas formas de funciones y tipos de datos, en esta tesis hemos presentado solo su aplicación a la optimización de funcionales de intercambio GGA a conjuntos de datos que consten de energías de 1-,2-, y 3-cuerpos de la expansión de muchos cuerpos del agua. Las Energías fueron optimizadas a un conjunto de datos de energías calculadas con el potencial MB-pol, que representa energías de agua al nivel de precisión de CCSD(T).

En una prueba del método de DPPS, hemos demostrado que, al optimizar el intercambio PBE a un conjunto de datos de energía calculadas con el funcional PW86R, la forma de intercambio de PW86R se reproduce precisamente. Esto confirma que DPPS funciona como queríamos en el caso de intercambio GGA optimizado a energías de agua. Después aplicamos DPPS a la optimización de la parte de intercambio GGA del funcional vdW-DF-cx para energías de 1-cuerpo, de 2-cuerpos, de 3-cuerpos, y de 2 + 3-cuerpos simultáneamente, sin y con constreñimientos Bayesianos. Los funcionales optimizados, sin y con constreñimientos, redujeron los errores en las energías de los conjuntos de datos. Aunque los resultados de los funciónales constreñidos resultan en más errores, produjeron formas funcionales más realistas. Considerando los errores en cada uno de los términos de energías de 1-,2-, y 3-cuerpos, los funcionales que son más precisos eran los que fueron optimizados a energías de 2-cuerpos (CUSTOM-2B) y a 2- y 3-cuerpos simultáneamente (CUSTOM-MB).

Debido a los limitados grados de libertad de un funcional de intercambio GGA, hemos encontrado dificultades en la optimización de este funcional a energías de 1-cuerpo simultáneamente con otros términos de energía. La optimización del término de energía de 1-cuerpo requiere cambios que no pueden estar reflejados en el intercambio GGA, lo que resulta en cambios no naturales en la forma funcional para compensar. Estos cambios están en conflicto con los necesarios para corregir energías de 2- y 3-cuerpos, y hacen imposible obtener correcciones significantes a las energías de 1-cuerpo simultáneamente con estos términos a través de la optimización del intercambio de GGA. Por lo tanto, para poder ajustar las energías de 1-cuerpo calculados con los funcionales CUSTOM-2B y CUSTOM-3B, añadimos un término de corrección a la energía total.

Antes de comprobar los funcionales optimizados con simulaciones de dinámica molecular ab initio (AIMD), hicimos varias simulaciones para encontrar para qué parámetros conseguimos resultados que converjan. Sin embargo, las propiedades obtenidas de las trayectorias de las simulaciones de AIMD tienen fluctuaciones estadísticas grandes. Desafortunadamente, las simulaciones de AIMD son demasiado caras para poder hacerlas durante tiempo suficiente para reducir los errores estadísticos significantemente. Para las fluctuaciones estadísticas grandes, la convergencia de unas propiedades no estaba clara con los resultados de AIMD. Para resolver estas dudas, hemos desarrollado un método eficiente para comprobar la convergencia con los parámetros de AIMD, basado en ideas de la integración termodinámica, lo que llamamos el método de Integración Termodinámica (TI). Con esto método, hacen falta cálculos con solamente el 1 por ciento de la trayectoria de una simulación de AIMD, y por eso resulta bastante más eficiente. Además, dado que el método TI usa la misma trayectoria para estimar cambios con cada parámetro, las propiedades resultantes con este método tienen menos fluctuaciones que los resultados de AIMD. Después de aplicar el método TI a las mismas pruebas de convergencia que hicimos con AIMD, fuimos capaces de resolver dudas sobre la convergencia con unos parámetros y confirmar la convergencia con otros.

Después de las pruebas de convergencia, usamos los funcionales optimizados para calcular las propiedades de clústeres de agua y agua líquida, y comparamos los resultados con los de MB-pol. En una prueba de las energías de isómeros de hexámeros de agua, descubrimos que la inclusión de correcciones de las energías de 2-cuerpos en CUSTOM-2B es suficiente para conseguir el orden de estabilidad apropiado en los hexámeros. Además, los resultados mejoran con el funcional CUSTOM-MB. Además, los dos funcionales optimizados redujeron el error en las energías de un rango más grande de geometrías de agua en un estudio de clústeres aleatorios sacados de una simulación de AIMD. Los resultados de estos cálculos con clústeres también enseñan que la contribución más grande a la energía total es el término de energías de 1-cuerpo.

Considerando las diferencias entre el funcional original vdW-DF-cx, y las versiones optimizadas con DPPS, los resultados de simulaciones de AIMD de agua liquida sugieren que la combinación de correcciones en las energías de 1-,2-, y 3-cuerpos es necesaria para predecir mejor la estructura del agua líquida. Mientras las correcciones de energías de 1-cuerpo no resultaron en cambios significantes

en la estructuración del agua líquida, cambiaron la distribución de geometrías de monómeros, alterando las distancias O-H y los ángulos H-O-H intra-molecular. El cambio en la distribución de geometrías de los monómeros mueve ligeramente la distribución de primeros vecinos de distancias más largas, debido al efecto de anticorrelación. Del mismo modo, el funcional CUSTOM-2B-D1, que incluye correcciones a energías de 2-cuerpos, no hace cambios en la estructuración del líquido pero hizo cambios en la distancia a primero vecinos, lo que lo pone más de acuerdo con MB-pol. El cambio más grande en la estructuración ocurrió con la inclusión de correcciones las energías de 3-cuerpos en el funcional CUSTOM-MB-D1, que hace la estructuración más suave, lo que pone los resultados aún más cerca de los resultados de MB-pol.

Appendix A

Constraints for the exchange-correlation energy

Here the important known exact constraints on the exchange-correlation energy are briefly explained. It has been suggested by Perdew *et al*. [101] that in the construction of exchange-correlation density functionals, the traditional non-empirical approach of construction using the satisfaction of constraints is the best approach rather than using semiempircal fitting, in which functionals are fit to selected data from experiment or from high accuracy *ab initio* calculations. In this thesis a method for functional development which combines both of these approaches is presented. The constraints discussed here were considered in the parameterization of a new, optimized exchange-correlation functional which is discussed in Chapter 4. These constraints are:

- *Self-interaction error correction (exchange and correlation)*: This constraint, which involves the cancellation of the energy of interaction of an electron density with itself, was discussed in Chapter 2.
- *Sum rule (exchange and correlation)*: The sum rule for the exchange-correlation hole is shown in Eq. 2.25 of Chapter 2.
- *Size-consistency (exchange and correlation)*: Size-consistency [101] can be demonstrated with two non-interacting systems, for example systems A and B. If the systems are separated by such a large distance that they essentially share no electron density, then

$$E(A+B) = E(A) + E(B), \tag{A.1}$$

where $E(A+B)$ is energy of the combined systems, and $E(A)$ and $E(B)$ are the energies of the subsystems. All DFT semilocal functionals satisfy this basic constraint.

- *Spin-scaling relation (exchange)*: This relation states that

$$E_x[n_\uparrow, n_\downarrow] = \frac{1}{2}[E_x[2n_\uparrow] + E_x[2n_\downarrow]]. \tag{A.2}$$

This relation is satisfied by all known density functionals [101].

- *Uniform-density scaling relation (exchange)*: This relation states that [64]

$$E_x[n_\gamma] = \gamma\, E_x[n], \tag{A.3}$$

where $n_\gamma = \gamma^3 n(\gamma\ \mathbf{r})$ and γ is a scale factor which changes the length scale of the density while maintaining normalization. This is also satisfied by all density functionals.

- *The Lieb–Oxford lower bound (exchange)*: This is the constraint that

$$E_x[n_\downarrow, n_\uparrow] \geq E_{xc}[n_\downarrow, n_\uparrow] \geq 2.273\ E_x^{LSDA}[n/2, n/2] \tag{A.4}$$

for all possible spin densities [67].

- *High-density limit (correlation)*: Levy [63] showed that in the high density limit, the correlation energy scales to a constant, i.e.,

$$\lim_{\gamma\to\infty} E_c[n_\gamma] = const. \tag{A.5}$$

- *Electron-nucleus cusp condition (exchange and correlation)*: This condition states that the exchange-correlation potential $V_{xc}^\sigma = \delta E_{xc}/\delta n(\mathbf{r}, \sigma)$ should be finite at the nuclear cusp of the electron-density [93].

References

[1] (2013). *CRC Handbook of Chemistry and Physics*. CRC Press, Raton, FL.

[2] Adamo, C. and Barone, V. (1999). *J. Chem. Phys.*, 110:6158.

[3] Allen, M. P. and Tildesley, D. J. (1987). *Computer Simulation of Liquids*. Oxford Univ. Press, Oxford.

[4] Babin, V., Leforestier, C., and Paesani, F. (2013). *J. Chem. Theory Comput.*, 9:5395.

[5] Babin, V., Medders, G. R., and Paesani, F. (2012). *J. Phys. Chem. Lett.*, 3:3765.

[6] Babin, V., Medders, G. R., and Paesani, F. (2014). *J. Chem. Theory Comput.*, 10:1599.

[7] Babin, V. and Paesani, F. (2013). *Chem. Phys. Lett.*, 580:1.

[8] Bachelet, G. B., Hamann, D. R., and Schlüter, M. (1982). Pseudopotentials that work: From h to pu. *Phys. Rev. B*, 26:4199–4228.

[9] Bates, D. M. and Tschumper, G. S. (2009). *J. Phys. Chem. A*, 113:3555.

[10] Becke, A. D. (1988). Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098–3100.

[11] Becke, A. D. (1993). Density functional thermochemistry iii the role of exact exchange. *J. Chem. Phys.*, 98:5648.

[12] Ben, M. D., Schönherr, M., Hutter, J., and VandeVondele, J. (2013). Bulk liquid water at ambient temperature and pressure from mp2 theory. *The Journal of Physical Chemistry Letters*, 4(21):3753–3759.

[13] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York.

[14] Berland, K. and Hyldgaard, P. (2014). Exchange functional that tests the robustness of the plasmon description of the van der waals density functional. *Phys. Rev. B*, 89:035412.

[15] Bohm, D. and Pines, D. (1951). *Phys. Rev.*, 82:625.

[16] Born, M. and Oppenheimer, J. R. (1927). Zur quantentheorie der molekeln. *Ann. Phys.*, 389:457–484.

[17] Ceperley, D. M. and Alder, B. J. (1980). Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.*, 45:566.

[18] Chen, B., Ivanov, I., Klein, M. L., and Parrinello, M. (2003). *Phys. Rev. Lett.*, 91:215503.

[19] Cooper, V. R. (2010). Van der waals density functional: An appropriate exchange functional. *Phys. Rev. B*, 81:161104.

[20] Corsetti, F., Artacho, E., Soler, J. M., Alexandre, S. S., and Fernández-Serra, M.-V. (2013a). Room temperature compressibility and diffusivity of liquid water from first principles. *The Journal of Chemical Physics*, 139(19).

[21] Corsetti, F., Fernández-Serra, M., Soler, J. M., and Artacho, E. (2013b). *J. Phys.: Condens. Matter*, 25:435504.

[22] Cui, J., Liu, H., and Jordan, K. D. (2006). *J. Phys. Chem. B*, 110:18872–18878.

[23] Dahlke, E. E., Olson, R. M., Leverentz, H. R., and Truhlar, D. G. (2008). *J. Phys. Chem. A*, 112:3976.

[24] Davidson, E. R. and Feller, D. (1986). Basis set selection for molecular calculations. *Chemical Reviews*, 86(4):681–696.

[25] de la Peña, L. H. and Kusalik, P. G. (2006). Quantum effects in liquid water and ice: Model dependence. *J. Chem. Phys.*, 125(5).

[26] Dion, M., Rydberg, H., Schröder, E., Langreth, D. C., and Lundqvist, B. I. (2004). Van der waals density functional for general geometries. *Phys. Rev. Lett.*, 92:246401.

[27] Fermi, E. (1934). *Nuovo Cimento*, 11:157.

[28] Feynman, R. P. (1939). Forces in molecules. *Phys. Rev.*, 56:340.

[29] Fiolhais, E. C., Nogueira, F., and Marques, M. (2003). *A Primer in Density Functional Theory*. Springer-Verlag, Berlin and Heidelberg.

[30] Fritsch, S., Potestio, R., Donadio, D., and Kremer, K. (2014). Nuclear quantum effects in water: A multiscale study. *Journal of Chemical Theory and Computation*, 10(2):816–824.

[31] Gillan, M. J., Manby, F. R., Towler, M. D., and Alfè, D. (2012). *J. Chem. Phys.*, 136:244105.

[32] Griffiths, D. J. (2005). *Introduction to Quantum Mechanics, 2nd Edition*. Pearson Prentice Hall, New Jersey.

[33] Grossman, J. C., Schwegler, E., Draeger, E. W., Gygi, F., and Galli, G. (2004). *J. Chem. Phys.*, 120:300.

[34] Guidon, M., Schiffmann, F., Hutter, J., and VandeVondele, J. (2008). Ab initio molecular dynamics using hybrid density functionals. *J. Chem. Phys.*, 128(21).

[35] Góra, U., Podeszwa, R., Cencek, W., and Szalewicz, K. (2011). *J. Chem. Phys.*, 135:224102.

[36] Habershon, S., Markland, T. E., and Manolopoulos, D. E. (2009). Competing quantum effects in the dynamics of a flexible water model. *J. Chem. Phys*, 131(2).

[37] Hamann, D. R., Schlüter, M., and Chiang, C. (1979). Norm-conserving pseudopotentials. *Phys. Rev. Lett.*, 43:1494–1497.

[38] Hammer, B., Hansen, L. B., and Nørskov, J. K. (1999). Improved adsorption energetics within density-functional theory using revised perdew-burke-ernzerhof functionals. *Phys. Rev. B*, 59:7413–7421.

[39] Harris, J. (1985). Simplified method for calculating the energy of weakly interacting fragments. *Phys. Rev. B*, 31:1770–1779.

[40] Hellmann, H. (1937). *Einfuehrung in die Quantenchemie*. Franz Deuticke, Leipzig.

[41] Hermann, A., Krawczyk, R., Lein, M., Schwerdtfeger, P., Hamilton, I., and Stewart, J. J. P. (2007). *Phys. Rev. A*, 76:013202.

[42] Hetsenyi, B., Angelis, F. D., Giannozzi, P., and Car, R. (2004). *J. Chem. Phys.*, 120:8632.

[43] Hodges, M. P., Stone, A. J., and Xantheas, S. S. (1997). *J. Chem. Phys. A.*, 101:9163–9168.

[44] Hohenberg, P. and Kohn, W. (1964). Inhomogeneous electron gas. *Phys. Rev.*, 136:B864.

[45] Hura, G., Russo, D., Glaeser, R. M., Head-Gordon, T., Krack, M., and Parrinello, M. (2003). *Phys. Chem. Chem. Phys.*, 5:1981.

[46] Jr., R. A. D., Santra, B., Li, Z., Wu, X., and Car, R. (2014). *J. Chem. Phys.*, 141:084502.

[47] Kerker, G. P. (1980). Non-singular atomic pseudopotentials for solid state applications. *Journal of Physics C: Solid State Physics*, 13(9):L189.

[48] Khaliullin, R., Cobar, E., Lochan, R., Bell, A., and Head-Gordon, M. (2012). *Phys. Chem. Chem. Phys.*, 14:15328–15339.

[49] Kim, K. and Jordan, K. D. (1994). Comparison of density functional and mp2 calculations on the water monomer and dimer. *The Journal of Physical Chemistry*, 98(40):10089–10094.

[50] Kim, K., Jordan, K. D., and Zwier, T. S. (1994). *J. Am. Chem. Soc.*, 116:11568.

[51] Kleinman, L. and Bylander, D. M. (1982). Efficacious form for model pseudopotentials. *Phys. Rev. Lett.*, 48:1425–1428.

[52] Klimes, J., Bowler, D. R., and Michaelides, A. (2010). Chemical accuracy for the van der waals density functional. *J. Phys.: Condens. Matter*, 22:022201.

[53] Klimeš, J. and Michaelides, A. (2012). Perspective: Advances and challenges in treating van der waals dispersion forces in density functional theory. *The Journal of Chemical Physics*, 137(12):120901.

[54] Kohn, W. (1987). Highlights of condensed-matter theory. proc. of the international school of physics “enrico fermi” course 89. *Acta Polym.*, 38:1521–4044.

[55] Kristyán, S. and Pulay, P. (1994). *Chem. Phys. Lett.*, 229:175.

[56] Kryachko, E. S. (1999). *Chem. Phys. Lett.*, 314:353.

[57] Kuharski, R. A. and Rossky, P. J. (1985). *J. Chem. Phys.*, 82:5164.

[58] Kuhne, T. D., Krack, M., and Parrinello, M. (2009). *J. Chem. Theory Comput.*, 5:235.

[59] Kumar, R., Schmidt, J. R., and Skinner, J. L. (2007). *J. Chem. Phys.*, 126:204107.

[60] Langreth, D. C. and Vosko, S. (1990). *Adv. Quantum. Chem.*, 21:175.

[61] Lee, C., Yang, W., and Parr, R. G. (1988). Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789.

[62] Lee, K., Murray, E. D., Kong, L., Lundqvist, B. I., and Langreth, D. C. (2010). Higher-accuracy van der waals density functional. *Phys. Rev. B*, 82:081101(R).

[63] Levy, M. (1989). *J. Quantum Chem. Quantum Chem. Symp*, 23:617.

[64] Levy, M. and Perdew, J. P. (1985). *Density Functional Methods in Physics*. Plenum, New York.

[65] Libnau, F. O., Toft, J., Christy, A. A., and Kvalheim, O. M. (1994). *J. Am. Chem. Soc.*, 116:8311.

[66] Libowitzky, E. (1999). *Monatsh. Chem.*, 130:1047.

[67] Lieb, E. H. and Oxford, S. (1981). *Int. J. Quantum Chem.*, 19:427.

[68] Lin, I.-C., Seitsonen, A. P., Coutinho-Neto, M. D., Tavernelli, I., and Rothlisberger, U. (2009). *J. Phys. Chem.*, 113:1127.

[69] Lin., J. S., Qteish, A., Payne, M. C., and Heine, V. (1993). Optimized and transferable nonlocal separable *ab initio* pseudopotentials. *Phys. Rev. B*, 47:4174–4180.

[70] Lobaugh, J. and Voth, G. A. (1997). A quantum model for water: Equilibrium and dynamical properties. *J. Chem. Phys.*, 106(6):2400–2410.

[71] MacDonald, J. K. L. (1933). Successive approximations by the rayleigh-ritz variation method. *Phys. Rev.*, 43:830–833.

[72] Martin, R. M. (2004). *Electronic Structure : Basic Theory and Practical Methods*. Cambridge University Press, Cambridge.

[73] Martin, R. M. (2008). *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, Cambridge.

[74] Mattsson, A. E. and Armiento, R. (2009). Implementing and testing the am05 spin density functional. *Phys. Rev. B*, 79:155101.

[75] Mayer, J. E. and Mayer, M. G. (1940). *Statistical Mechanics*. John Wiley and Sons Inc, Hoboken, NJ.

[76] Medders, G., Babin, V., and Paesani, F. (2014). *J. Chem. Theory Comput.*, 10:2906.

[77] Medders, G. R., Babin, V., and Paesani, F. (2013). *J. Chem. Theory Comput.*, 9:1103.

[78] Medders, G. R., Götz, A. W., Morales, M. A., Bajaj, P., and Paesani, F. (2015). On the representation of many-body interactions in water. *J. Chem. Phys.*, 143(10).

[79] Miller, T. F. and Manolopoulos, D. E. (2005). Quantum diffusion in liquid water from ring polymer molecular dynamic. *J. Chem. Phys.*, 123(15).

[80] Miller, W. H. (2005). Quantum dynamics of complex molecular systems. *Proc. Nat. Acad. Sci. USA*, 102(19):6660–6664.

[81] Minceva-Sukarova, B., Sherman, W. F., and Wilkinson, G. R. (1984). *J. Phys. C*, 17:5833.

[82] Møller, C. and Plesset, M. S. (1934). Note on an approximation treatment for many-electron systems. *Phys. Rev.*, 46:618–622.

[83] Morales, M. A., Gergely, J. R., McMinis, J., McMahon, J. M., Kim, J., and Ceperley, D. M. (2014). *J. Chem. Theory Comput.*, 10:2355–2362.

[84] Morrone, J. A. and Car, R. (2008). *Phys. Rev. Lett.*, 101:017801.

[85] Murray, E. D., Lee, K., and Langreth, D. C. (2009). *J. Chem. Theory Comput.*, 5:2754.

[86] Møgelhøj, A., Kelkkanen, A. K., Wikfeldt, K. T., Schiøtz, J., Mortensen, J. J., Pettersson, L. G. M., Lundqvist, B. I., Jacobsen, K. W., Nilsson, A., and Nørskov, J. K. (2011). Ab initio van der waals interactions in simulations of water alter structure from mainly tetrahedral to high-density-like. *The Journal of Physical Chemistry B*, 115(48):14149–14160. PMID: 21806000.

[87] Møller, C. and Plesset, M. S. (1934). *Phys. Rev.*, 46:618.

[88] Odashima, M. M., Capelle, K., and Trickey, S. B. (2009). *J. Chem. Theory Comput.*, 5:798.

[89] Paesani, F., Iuchi, S., and Voth, G. A. J. (2007). *J. Chem. Phys.*, 127:074506.

[90] Paesani, F. and Voth, G. A. (2009). The properties of water: Insights from quantum simulations†. *J. Phys. Chem. B*, 113(17):5702–5719. PMID: 19385690.

[91] Paesani, F., Zhang, W., Case., D. A., Cheatham, T. E., and Voth, G. A. (2006). An accurate and simple quantum model for liquid water. *J. Chem. Phys.*, 125(18).

[92] Pamuk, B., Soler, J. M., Ramírez, R., Herrero, C. P., Stephens, P. W., Allen, P. B., and Fernández-Serra, M.-V. (2012). Anomalous nuclear quantum effects in ice. *Phys. Rev. Lett.*, 108:193003.

[93] Pan, X.-Y. and Sahni, V. (2003). *Phys. Rev. A*, 67:012501.

[94] Patridge, H. and Schwenke, D. W. (1997). *J. Chem. Phys.*, 106:4618.

[95] Pedroza, L. S., da Silva, A. J. R., and Capelle, K. (2009). *Phys. Rev. B*, 79:201106(R).

[96] Pedulla, J. M. (1996). *J. Chem. Phys.*, 105:11091.

[97] Perdew, J. P. and Burke, K. (1996). Comparison shopping for a gradient-corrected density functional. *Int. J. Quant. Chem.*, 57:309–319.

[98] Perdew, J. P., Burke, K., and Ernzerhof, M. (1996a). Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868.

[99] Perdew, J. P., Ernzerhof, M., and Burke, K. (1996b). *J. Chem. Phys.*, 105:9982.

[100] Perdew, J. P., Ruzsinszky, A., Csonka, G. I., Vydrov, O. A., Scuseria, G. E., Constantin, L. A., Zhou, X., and Burke, K. (2008). Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.*, 100:136406.

[101] Perdew, J. P., Ruzsinszky, A., Tao, J., Staroverov, Viktor, N., Scuseria, G. E., and Csonka, G. I. (2005). Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *The Journal of Chemical Physics*, 123(6).

[102] Perdew, J. P. and Yue, W. (1986). Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Phys. Rev. B*, 33:8800–8802.

[103] Perdew, J. P. and Zunger, A. (1981). Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, 23:5048–5079.

[104] Perdew, P. and Wang, Y. (1994). *J. Chem. Phys.*, 100:1290.

[105] Ponder, J. W., Wu, C., Ren, P., Pande, V. S., J. D. Chodera, D. L. M., Schnieders, M. J., Haque, I., Lambrecht, D. S., R. A. DiStasio, J., Head-Gordon, M., Clark, G. N. I., Johnson, M. E., and Head-Gordon, T. (2010). *J. Phys. Chem.*, 114:2549.

[106] Prendergast, D. and Galli, G. (2006). *Phys. Rev. Lett.*, 96:215502.

[107] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes*. Cambridge University Press, Cambridge.

[108] Pérez-Jordá, J. M. and Beck, A. D. (1995). *PChem. Phys. Lett.*, 233:134.

[109] Raghavachari, K., Trucks, G. W., Pople, J. A., and Head-Gordon, M. (1989). *J. Chem. Theory Comput.*, 157:479.

[110] Rappe, A. M., Rabe, K. M., Kaxiras, E., and Joannopoulos, J. D. (1990). Optimized pseudopotentials. *Phys. Rev. B*, 41:1227–1230.

[111] Roman-Perez, G. and Soler, J. M. (2009). Efficient implementation of a van der waals density functional: Application to double-wall carbon nanotubes. *Phys. Rev. Lett.*, 103:096102.

[112] Russo, J. and Tanaka, H. (2014). *Nature Comm.*, 5:3556.

[113] Santra, B., Klimes, J., Alfe, D., Tkatchenko, A., Slater, B., Michaelides, A., Car, R., and Scheffler., M. (2011). *Phys. Rev. Lett.*, 107:185701.

[114] Santra, B., Michaelides, A., Fuchs, M., Tkatchenko, A., Filippi, C., and Scheffler, M. (2008). *J. Chem. Phys.*, 129:194111.

[115] Schmidt, J., VandeVondele, J., Kuo, I.-F. W., Sebastiani, D., Siepmann, J. I., Hutter, J., and Mundy, C. J. (2009). *Phys. Chem. B*, 113:11959.

[116] Schwegler, E., Grossman, J. C., Gygi, F., and Galli, G. (2004). *J. Chem. Phys.*, 121:5400.

[117] Skinner, L. B. and Benmore, C. J. (2012). *Sci. Rep.*, 2:505.

[118] Soler, J. M., Artacho, E., Gale, J. D., García, A., Junquera, J., Ordejón, P., and Sánchez-Portal, D. (2002). The SIESTA method for ab initio order-N materials simulation. *J. Phys.: Condens. Matter*, 14:2745–2779.

[119] Soper, A. K. (2007). *J. Phys.: Condens. Matter*, 19:335206.

[120] Spura, T., John, C., Habershon, S., and Kühne, T. D. (2015). Nuclear quantum effects in liquid water from path-integral simulations using an ab initio force-matching approach. *Mol. Phys.*, 113(8):808–822.

[121] Staroverov, V. N., Scuseria, G. E., Tao, J., and Perdew, J. P. (2004). *Phys. Rev. B*, 69:075102.

[122] Temelso, B., Archer, K. A., and Shields, G. C. (2011). *J. Phys. Chem. A*, 115:12034–12046.

[123] Tkatchenko, A. and Scheffler, M. (2009). *Phys. Rev. Lett.*, 102:073005.

[124] Troullier, N. and Martins, J. L. (1991). Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B*, 43:1993–2006.

[125] Vanderbilt, D. (1985). Optimally smooth norm-conserving pseudopotentials. *Phys. Rev. B*, 32:8412–8415.

[126] Vydrov, O. A., Scuseria, G. E., Perdew, J. P., Ruzsinszky, A., and Csonka, G. I. (2006). Scaling down the perdew-zunger self-interaction correction in many-electron regions. *J. Chem. Phys.*, 124(9).

[127] Vydrov, O. A. and Voorhis, T. V. (2010). Nonlocal van der waals density functional: The simpler the better. *J. Chem. Phys.*, 133:244103.

[128] Wallqvist, A. and Berne, B. J. (1985). *Chem. Phys. Lett.*, 117:214.

[129] Wang, J., Román-Pérez, G., Soler, J. M., Artacho, E., and Fernández-Serra, M.-V. (2011). *J. Chem. Phys.*, 134:024516.

[130] Wang, Y., Babin, V., Bowman, J. M., and Paesani, F. (2012). *J. Am. Chem. Soc.*, 134:11116.

[131] Woutersen, S., Emmerichs, U., Nienhuys, H. K., and Bakker, H. J. (1998). *Phys. Rev. Lett.*, 81:1105.

[132] Wu, Z. and Cohen, R. E. (2006). More accurate generalized gradient approximation for solids. *Phys. Rev. B*, 73:235116.

[133] Xantheas, S. S. (1994). *J. Chem. Phys.*, 100:7523–7534.

[134] Xantheas, S. S. (2000). *Chem. Phys.*, 258:225–231.

[135] Zhang, Y. and Yang, W. (1998). Comment on “generalized gardient approximation made simple”. *Phys. Rev. Lett.*, 80:890.

[136] Zhao, Y. and Truhlar, D. G. (2005). *J. Chem. Theory Comput.*, 1:415.

[137] Zupan, A., Burke, K., Ernzerhof, M., and Perdew, J. P. (1997). Distributions and averages of electron density parameters: Explaining the effects of gradient corrections. *J. Chem. Phys.*, 106:10184–10193.

.